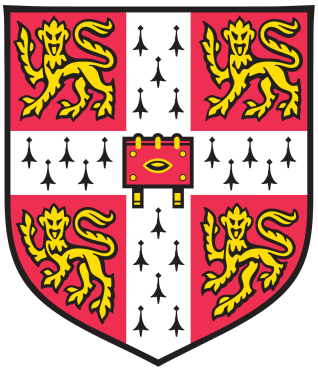
Application of Deep Learning to Brain Connectivity Classification in Large MRI Datasets

Matthew Leming

A thesis presented for the degree of
Doctor of Philosophy



Department of Psychiatry
University of Cambridge
United Kingdom

June 2020

Acknowledgements

I would like to thank those individuals that contributed to the published and submitted forms of this research, which include Shayanti Chattopadhyay, Li Su, Juan Manuel-Gorríz, Simon Baron-Cohen, and, of course, my supervisor John Suckling. I would also like to thank those individuals that contributed through either informal advice and conversations to the development of this work or whose data releases were essential to carrying out several studies, including Sarah Morgan, Richard Bethlehem, Varun Warrier, Lena Dorfschmidt, and Ed Bullmore. And I would like to thank Luca Villa and Ayan Mandal, with whom I have had frequent conversations about neuroscience in general, which helped with my overall understanding of the field.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 60,000 words excluding appendices, bibliography, footnotes, tables, equations, and has fewer than 150 figures.

Matthew Leming, June 2020

Abstract

Title: Application of Deep Learning to Brain Connectivity Classification in Large MRI Datasets

The use of machine learning for whole-brain classification of magnetic resonance imaging (MRI) data is of clear interest, both for understanding phenotypic differences in brain structure and function and for diagnostic applications. Developments of deep learning models in the past decade have revolutionized photographic image and speech recognition, bringing promise to do the same to other fields of science. However, there are many practical and theoretical challenges in the translation of such methods to the unique context of MRIs of the brain. This thesis presents a theoretical underpinning for whole-brain classification of extremely large datasets of multi-site MRIs, including machine learning model architecture, dataset curation methods, machine learning visualization methods, encoding of MRI data, and feature extraction. To replicate large sample sizes typically applied to deep learning models, a dataset of over 50,000 functional and structural MRIs was amassed from nine different databases, and the undertaken analyses were conducted on three covariates commonly found across these collections: sex, resting state/task, and autism spectrum disorder. I find that deep learning is not only a method that has promise for clinical application in the future, but also a powerful statistical tool for analyzing complex, nonlinear relationships in brain data where conventional statistics may fail. However, results are also dependent on factors such as dataset imbalances, confounding factors such as motion and head size, selected methods of encoding MRI data, variability of machine learning models and selected methods of visualizing the machine learning results. In this thesis, I present the following methodological innovations: (1) a method of balancing datasets as a means of regressing out measurable confounding factors; (2) a means of removing spatial biases from deep learning visualization methods; (3) methods of encoding functional and structural datasets as connectivity matrices; (4) the use of ensemble models and convolutional neural network architectures to improve classification accuracy and consistency; (5) adaptation of deep learning visualization methods to study brain connections utilized in the classification process. Additionally, I discuss interpretations, limitations, and future directions of this research.

Matthew Leming, June 2020

List of Publications

- Leming, M., Su, L., Chattopadhyay, S., Suckling, J. "Normative Pathways in the Functional Connectome." NeuroImage 184(1): 317-334. 2019.
- Leming, M. and Suckling, J. "Deep Learning on Brain Images in Autism: What Do Large Samples Reveal of Its Complexity?" Proceedings of the 8th International Work-Conference on the Interplay Between Natural and Artificial Computation, Part I (Springer, 2019): 389–402.
- Leming, M., Manuel-Gorríz, J., and Suckling, J. "Ensemble deep learning on large, mixed-site fMRI datasets in autism and other tasks." International Journal of Neural Systems 30(7): 2050012-1–16. 2020.
- Leming, M. and Suckling, J. "Stochastic encoding of graphs in deep learning allows for complex analysis of gender classification in resting-state and task functional brain networks from the UK Biobank." (Submitted to IEEE Trans. on Pattern Analysis), 2020. (https://arxiv.org/abs/2002.10936)
- Leming, M., Baron-Cohen, S., and Suckling, J. "Single-participant structural connectivity matrices lead to greater accuracy in classification of participants than function in autism in MRI." (Submitted to IEEE Trans. on Med. Imaging). 2020. (https://arxiv.org/abs/2005.08035)

Contents

1 Introduction **1**

1.1 MRI 1

1.1.1 Structural MRI 1

1.1.2 Functional MRI 2

1.2 Graph theory 3

1.3 Connectivity 4

1.3.1 Functional connectivity 4

1.3.2 Structural connectivity 6

1.3.3 Multi-slice connectomes 7

1.3.4 Identification of brain networks 8

1.4 Machine learning 10

1.4.1 Overview 10

1.4.2 Training 10

1.4.3 Overview of machine learning methods 11

1.4.4 Convolutional neural networks to classify graphs 12

1.4.5 Explainable AI and the black box problem 13

1.4.6 Methods of visualizing deep neural networks 14
1.4.7 Machine learning in medical imaging of the brain 17
1.5 Big data in medical imaging 19
1.5.1 Overview 19
1.5.2 Role in machine learning 19
1.6 Connectivity in different cohorts 20
1.6.1 Task fMRI 21
1.6.2 Sex 21
1.6.3 Autism 25
1.6.4 Depression 27
1.7 Research Aims 27
1.8 Thesis outline 28
1.9 Motivation and Themes 29

2 Amassing and processing large datasets 31

2.1 Data acquisition 31
2.1.1 Dataset descriptions and labeling 32
2.1.2 Distribution of data quality 35
2.2 FMRI signal processing toolbox 35
2.3 Dataset counts 36
2.4 Deep learning model 37
2.4.1 Implementation 37

2.4.2 Implementation of visualization methods 38
2.5 Use of Connectivity 39
2.6 Practical limitations 40
2.7 Hyperparameter tuning 41
2.8 Note about AUROC and accuracy 43

3 Brain connectivity analysis for mental conditions 45

3.1 Normative pathways 45
3.1.1 Introduction 46
3.1.2 Methods 49
3.1.3 Results 59
3.1.4 Conclusion 76
3.2 A novel structural connectivity metric 77
3.2.1 Introduction 77
3.2.2 Methods 78
3.2.3 Results 80
3.2.4 Discussion 81

4 Ensemble CNNs for connectivity classification 83

4.1 Introduction 83
4.2 Methods 85
4.2.1 Datasets and preprocessing 85
4.2.2 Neural network model and training 86

4.2.3 Set division . . . 88
4.2.4 Test set evaluation . . . 88
4.2.5 Experiments . . . 89
4.3 Results . . . 90
4.3.1 Autism vs TD controls . . . 90
4.3.2 Sex . . . 92
4.3.3 Rest vs task . . . 92
4.3.4 Ensemble model limits . . . 93
4.4 Discussion . . . 93
4.5 Conclusion . . . 95

5 Activation maximization **97**
5.1 Introduction . . . 97
5.2 Methods . . . 98
5.3 Results . . . 99
5.3.1 Autism vs TD controls . . . 100
5.3.2 Sex . . . 101
5.3.3 Rest vs task . . . 101
5.4 Discussion . . . 102

6 Multivariate class balancing **105**
6.1 Introduction . . . 106
6.2 Methods . . . 107

6.2.1 Data pre-processing . . . 107
6.2.2 Multivariate class balancing . . . 108
6.2.3 Formalization of multivariate class balancing problem . . . 108
6.2.4 Balancing algorithm . . . 109
6.3 Results . . . 110

7 Salience in brain connectomes 111

7.1 Introduction . . . 111
7.1.1 Network brain function across the sexes . . . 113
7.2 Methods . . . 116
7.2.1 Machine learning . . . 116
7.2.2 Visualization of machine learning results . . . 117
7.3 Results . . . 121
7.3.1 Machine learning . . . 121
7.3.2 Visualization of machine learning results . . . 121
7.4 Discussion . . . 128
7.4.1 Deep learning model . . . 128
7.4.2 Neuroscientific findings of CAMs from vertical filters (from Chapter 4) 130
7.4.3 Neuroscientific interpretations of CAMs from sex classification in UK BioBank . . . 131
7.5 Conclusion . . . 133

8 Structure/function encoding in autism 135

8.1 Introduction 135
8.1.1 Studies of the structure-function relationship in the brain 137
8.1.2 Experiments 139
8.2 Methods 139
8.2.1 Dataset 139
8.2.2 Pre-processing and feature extraction 140
8.2.3 Machine learning model and training 140
8.2.4 Class activation map analysis 142
8.3 Results 142
8.3.1 Training 142
8.3.2 Class activation map analysis 144
8.4 Discussion 146
8.5 Conclusion 149

9 General discussion **153**

9.1 Summary 153
9.2 Big versus small datasets 155
9.3 Psychiatric diagnosis in machine learning 156
9.4 Class balancing techniques 156
9.5 Comparison of machine learning encoding methods 158
9.6 Interpretation of visualization methods 160
9.7 Shortcomings of salience detection methods 161

9.8 Failed research directions 161

9.9 Future directions 163

9.10 Conclusion 165

Chapter 1

Introduction

1.1 MRI

1.1.1 Structural MRI

The phenomenon of nuclear magnetic resonance (NMR) was originally described in the first half of the 20th century (Purcell et al., 1946; Bloch et al., 1946). NMR refers to the physical observation that atomic nuclei, when held in a strong magnetic field, may be perturbed by a weaker magnetic field and, when released, send out an electromagnetic pulse as the nucleus returns to its original alignment. This outgoing pulse is proportional to the strength of the applied magnetic field, which is the property that MRI takes advantage of to acquire 3D images (Damadian, 1971; Lauterbur, 1973; Mansfield and Maudsley, 1977). By applying and varying magnetic field strengths in unique intensity ranges in three dimensions (thereby assuring that only localized protons with a certain field strength applied will reply to a magnetic pulse) and detecting the output nuclear magnetic resonance strengths, a 3D k-space varying across spatial frequencies can be filled (or, more commonly, many slices of 2D k-spaces that compose a 3D space). By the application of a Fourier transform, a 3D representation of water density in tissue can be derived.

3D MR images can take different forms by varying the repetition time (TR) (or the time between successive pulse sequences applied to a particular slice) and the echo time (TE) (or the time between the delivery of the magnetic pulse and the reception of the echo signal). The most common of these forms are T1- and T2-weighted images. T1, the longitudinal

relaxation time, refers to the time required for shifted electrons to return to equilibrium with the strong magnetic field after the pulse is removed, and can be emphasized in MRIs with short echo and repetition times. T2, the transverse relaxation time, refers to the time taken for excited protons to lose phase with each other, measuring the coherence of nuclei spinning perpendicular to the main magnetic field, and can be obtained with long repetition and echo times. Practically, T1 is useful for viewing brain structure (with cerebrospinal fluid appearing darker and brain matter appearing brighter), while T2 is more often used to view lesions (with fluids appearing brighter). Other variations of 3D MRI can be obtained by varying TE and TR, such as fluid-attenuated inversion recovery (FLAIR) and proton density imaging, but T1- and T2-weighted MRIs are the most commonly used.

The 1990s saw several extensions of MRI into four dimensions, through rapid acquisition of multiple 3D volumes with different settings, that produced a number of new innovations in the field. Most notably, diffusion-weighted imaging, which produces diffusion tensor imaging (Basser et al., 1994), has aided the study of white matter tracts in the brain, while functional MRI (Ogawa et al., 1990) produced a means of studying localized brain function.

1.1.2 Functional MRI

Functional MRI (fMRI) (Ogawa et al., 1990) builds on 3D MRI by taking advantage of the magnetic properties of hemoglobin (Pauling and Coryell, 1936) to show variation in brain metabolism across time. This is called the blood-oxygen level dependent (BOLD) signal, which is believed to be an indirect indicator of brain activity. The use of rapid 3D acquisition allows for the measurement of the BOLD signal within a given period of time. However, due to time constraints, these individual slices are often of lower resolution than typical structural MRI, and even with different optimizations applied, fMRI's temporal resolution, being limited by acquisition speeds and the haemodynamic response that composes the BOLD signal, is within the range of 1-2 seconds. This puts it at a disadvantage over methods of detecting brain activity, such as EEG, which has a higher temporal resolution, but it is the best means available of studying it in localized areas of the brain, even though the BOLD signal is considered an indirect indication of brain function.

FMRI has not yet found widespread clinical use, but is most often applied in psychological studies, typically either to test brain activation in different psychological tasks, or to compare resting-state brain activity in different populations. Task studies have found that different tasks characterize different localized activations and brain network patterns across different

populations, but the nature of the tasks, differences in population sample sizes and covariates, and vastly different preprocessing and analysis methods, often make it difficult to compare two different task fMRI studies. The use of resting-state fMRI alleviates the first of these issues, though had been found to have a higher variability than task fMRI when only a single task is considered (Elton and Gao, 2015).

FMRI is hugely affected by confounding factors, notably head motion, which affects the reading of the BOLD signal (Duncan and Northoff, 2013). This has particularly high potential to affect the outcomes of studies when motion is measurably different between two populations being studied, which is often the case. While physical constraints and sedatives may provide a means of reducing motion in the patient, these are either inconsistently applied, undesirable for a study, or ineffective at completely eliminating motion effects. A number of preprocessing steps have been proposed to regress out the effects of motion in the acquired data (Caballero-Gaudes and Reynolds, 2017), including registration of the 3D timepoints, head re-alignment (Goto et al., 2016), wavelet despiking (Patel et al., 2014), and more advanced methods (Kundu et al., 2012, 2013), but because the spin-echo effects of motion on water molecules are extremely difficult to model, none have been able to completely eliminate the effects of motion. While it may be difficult to fully regress, it is possible to detect the amount of motion by measuring the displacement of one 3D slice with its neighbor (Freire et al., 2002). A common practice in fMRI studies is to remove subjects with excessive measured head motion from the study and apply standard motion regression methods to the remaining data.

1.2 Graph theory

Graph theory is the study of nodes (or vertices) interconnected by edges that compose networks (or graphs). It is used to model many real-world systems, such as airline routes, road systems, and social networks. Graph theory has seen extensive development in the fields of computer science, statistics, and mathematics, and this wide development is often borrowed by other fields to analyze complex scientific data for which a graph representation can be found (Papo et al., 2014). Because of this, graph theory is a favorable means to model and analyze brain networks. This field is often called “connectivity”. For instance, two frequent applications of graph theory to brain connectivity are assessments of node centrality (Zuo et al., 2011; van den Heuvel and Sporns, 2013) (i.e., determining which nodes are “important” in a graph) and community partitioning (Sporns and Betzel, 2016)

(i.e., determining ways to separate networks into smaller subnetworks).

1.3 Connectivity

In brain connectivity, MRI datasets are reduced to a graph representation that is referred to as a "connectome". In this context, while brain "connections" may represent some type of physiological relationship between two areas of the brain, they may also represent measurements of functional or physiological similarity. Other than being a function that reduces an MRI dataset to a network representation, different methods of estimating brain connectivity may have little else in common. In this section, I review different types of connectivity and previous work on identifying subnetworks in brain connectomes.

1.3.1 Functional connectivity

Since its inception, many computational methods have been developed to analyze fMRI data. One such method, functional connectomics (Friston et al., 1993), reduces the dimensionality of fMRI datasets to graphs (or networks), comprising nodes, representing brain areas, connected by edges, that represent the relationships between the measured BOLD signals (usually reduced to a timeseries) in these localized areas of the brain. While this dimensionality reduction simplifies the data and does away with a large amount of signal, it does allow for the use of graph theoretical methods, which has been extensively developed in pure mathematics and computer science. Furthermore, depending on the method used to construct the connectome, it allows for the direct analysis of relationships between regions of the brain. This enables the study of brain networks and pathways.

Graph theory metrics, when applied to functional connectomes, can estimates the qualities of brain organization with measurements such as centrality (or "hubness") (Sporns et al., 2007; Joyce et al., 2010; Lohmann et al., 2010; Rubinov and Sporns, 2010; Tomasi and Volkow, 2010, 2011b; Zuo et al., 2011) and community structure (or "modularity") (Traag and Bruggeman, 2009; Mucha et al., 2010; Bassett et al., 2013; Sporns and Betzel, 2016). In general, the functional connectome is characterized by high complexity (Sporns et al., 2000; Sporns, 2006), high efficiency (Buzsaki et al., 2004), global and local synchronizability (Masuda and Aihara, 2004), and high levels of clustering with short path lengths (Hilgetag et al., 2000; Stephan et al., 2000; Bassett and Bullmore, 2006), indicating a small-world

architecture (Milgram, 1967; Watts and Strogatz, 1998). Functional connectomes are also unique to individuals; this is called the connectome "fingerprint" (Finn et al., 2015). Fox et al. (2005) posited that the brain is organized into anticorrelated functional networks distributed over a wide area.

Individual brain networks have been consistently discerned from functional connectivity – for instance, the default mode network (DMN) in resting-state – though there is some disagreement as to the makeup and discernment of these networks Stanley et al. (2013). Research in brain networks is often performed with methods other than functional connectivity, such as independent component analysis (ICA). Some brain parcellations model common a-priori networks rather than smaller, anatomical areas of the brain. Discrepancies in brain networks are exacerbated by technical factors such as the parcellation used to derive the network; function-structure couplings (i.e, whether certain functional areas appear consistently in a particular anatomical area); whether the subject was in task or resting state and, if task, which task was performed; and other factors (Power et al., 2011; Sung et al., 2018).

Types of functional connectivity

Given a 2D timeseries, composed of a 1D signal for each parcellated area, functional connectivity can be derived using one of the many time series comparison metrics available; the output of a connectome parcellation program with N areas and M timepoints is a timeseries of size $N \times M$. Using a timeseries comparison metric, this can be transformed into an $N \times N$ matrix. Due to its widespread use in other fields, the most commonly favored metric is Pearson correlation, though the use of partial correlation, average mutual information, coherence, multi-band wavelet correlation, and others, have been proposed, with each having its own features and limitations, such as linearity versus nonlinearity, regression of global signal, and widespread use that contributes to general understanding (Bastos and Schoffelen, 2016). The use of predictive metrics (such as Granger causality (Ding et al., 2006)) may also be applied, but this models a causal relationship between brain regions and is more often referred to as "effective connectivity", rather than functional connectivity (Friston, 1994; Kriston, 2011).

Different timeseries comparison metrics can lead to the expression of different properties and connectomes with very different topologies; furthermore, differences in the output domain of different timeseries metrics affect the analysis techniques. For instance, any type of correlation values are between -1 and 1; these negative correlations give rise to a problem

of interpretability that is not easy for neuroscientists to address and which has generated a wide debate in connectomics (Zhan et al., 2017). However, mutual information, another comparison technique, produced values that are above 0, effectively avoiding this issue.

The most common metric of comparison is Pearson's correlation, which captures the linear relationship between timeseries values. While it is quick to compute and, given its prevalence in statistics, easy to interpret, it also fails to adequately capture nonlinear relationships between timeseries. Mutual information is a common alternative that does capture nonlinear solutions; however, while there is a theoretical basis of mutual information in information theory, there is no standardized way to calculate it on real-world data, leading to a number of different implementations. Partial correlation is a metric that regresses out every other timeseries when comparing two timeseries. It is the inverse of the correlation matrix. A quality of partial correlation is that it effectively regresses the global signal, including unwanted fluctuations in the BOLD signal due to blood flow. However, given too many parcellation areas and too few timepoints, partial correlation may regress out too much of the signal for useful analysis.

These three metrics may also be repeated on the timeseries following a wavelet transform (Patel and Bullmore, 2016); wavelet transforms operate on the signal, producing a decomposition in different frequency bands (requiring information about the repetition time of the MRI when comparing across sites), and, for Q different frequency ranges, this allows for the transformation of an $N \times M$ timeseries into an $N \times N \times Q$ matrix. Often, the analysis of such wavelet correlation matrices are performed on just one of these $N \times N$ matrices, as it is more difficult to interpret a multi-slice matrix as a graph, though more advanced methods allow for the analysis of multi-slice connectomes (Bassett et al., 2013).

1.3.2 Structural connectivity

Tractography and diffusion-based methods

The most well-known structural connectivity methods are based on white matter tract tracing between regions (Basser et al., 1994). While this produces the same output format as functional connectivity – an undirected adjacency matrix representing a single subject – they are methodologically very different. Structural connectivity, in its most widely-used form, represents the integrity of white matter fiber tracts. Producing such representations requires diffusion-weighted imaging, as this encodes information about the Brownian motion of water

molecules, which is directionally restricted by white matter tracts, allowing for the structure of white matter to be inferred (Jones et al., 2013). The strength of the connection can be quantified in several ways, such as the probability of a connection, fiber length, fiber density, or fiber count (Jones et al., 2013; Ji et al., 2014).

Structural covariance

Structural covariance (Wright et al., 1999) is an alternative means of constructing structural brain networks, though it is applied to represent groups, rather than individual subjects. After an N-area parcellation to a group of M structural MRIs, some scalar measurement (such as cortical thickness (Gong et al., 2012)) is estimated for each subject within each parcellation, producing $N \times M$ values. Like functional timeseries, these values may then be correlated. This offers a description of how different brain measurements of brain structure may relate to one another across populations.

Structural covariance networks have similar properties to functional connectivity networks, such as small worldness, nonrandom clustering, and modularity. Such properties, however, can also be seen in many real-world networks (Alexander-Bloch et al., 2013a) and may be linked to the tendency of structures to commonly co-vary over short distances (Chen et al., 2008) and occasionally over long distances; for instance, symmetrical regions of the brain, though often spatially disparate, tend to co-vary (Mechelli et al., 2005). Typically, spatial proximity is an indicator of both higher structural covariance between two regions and correlated functional activity (Salvador et al., 2005; Honey et al., 2009; Alexander-Bloch et al., 2013b).

1.3.3 Multi-slice connectomes

Given the diverse means available of analyzing timeseries, some are able to produce a multi-slice functional connectome (i.e, a $K \times N \times N$ matrix, as opposed to a single-slice $N \times N$ matrix). Such connectomes play an important role in this thesis.

Research in the analysis of multi-slice matrices is more difficult due to the complexity of analysis and visualization of an added dimension, and the under-development of it in other fields compared to single-slice graphs. Because of the more diverse means of calculating timeseries comparisons, multislice analysis is more developed in functional connectivity than

in structural connectivity. Even with developments that show high potential for multi-slice matrices, such as multi-level wavelet correlation (Patel et al., 2014; Patel and Bullmore, 2016) and dynamic functional connectivity (Calhoun et al., 2014; Calhoun and Adali, 2016; Preti et al., 2017), a typical practice is often to average such matrices into a single one, or to select a single timeseries comparison, or to analyze timeseries comparisons separately Achard et al. (2006); Achard and Bullmore (2007); Mucha et al. (2010).

Wavelet correlation (Bullmore et al., 2004; Achard et al., 2006; Achard and Bullmore, 2007; Skidmore et al., 2011) is the correlation of different scales in the wavelet decomposition of an fMRI timeseries, which allows for the building of a multislice connectome representative of the whole timeseries. Though it is more common to use a wavelet decomposition (Bullmore et al., 2004; Zhang et al., 2016b) for preprocessing (Patel et al., 2014), they have been used to derive multislice connectomes by correlating regional timecourses in different frequencies (Achard et al., 2006; Thompson and Fransson, 2015); most often, however, after deriving such a multislice connectome, slices are analyzed independently (Berlingerio et al., 2011).

1.3.4 Identification of brain networks

The application of graph theory to MRI data has led to the identification of anatomical and functional brain networks. This can be applied to the context of structural connectivity, which most often refers to white matter networks, but is most often used in reference to functional connectivity, indicating groups of disparate brain areas that are functionally similar under certain tasks (Bressler and Menon, 2010). These networks have been characterized by general, global properties, quantified by graph theoretical measurements, as well as identification of subnetworks in functional MRI data that have been associated with different properties. While identification of these networks is subject to the selected psychological task, brain parcellation, and analysis methods, a few common networks are consistently identified.

Common brain networks

The identification of brain networks in functional data in the literature is often dependent on the aim of the study and datasets analyzed. For instance, Sung et al. (2018) identified 111 brain networks related to psychometric parameters, while Smith et al. (2009) identified just 10. Greene et al. (2018) listed 10 “canonical” networks: the medial frontal, frontoparietal,

default mode, motor cortex, visual A, visual B, visual association, salience, subcortical, and cerebellum. Generally, the networks discussed the most in literature of resting-state and task fMRI are the default mode (Raichle et al., 2001), executive control, salience (Seeley et al., 2007), dorsal attention, and ventral attention networks (Vossel et al., 2014).

Issues in quantifying brain networks

Atlas-based parcellations are used to derive a brain network using predefined ROIs, defined either by anatomical regions, common functional areas, or randomly. Brain networks are affected by the selection of the parcellation atlas to a large extent (Yao et al., 2015). This is true in a trivial sense, as finer parcellations are able to quantify finer sections of the brain and lead to more detailed networks, but atlases also affect measurements derived from such networks; Wang et al. (2009) found that small worldness is substantially different depending on the selected parcellation atlas, and Zalesky et al. (2010) found that node scale in randomly parcellated structural connectomes affected global connectivity measurements substantially. Brain network estimation can be affected as well by spatial variability of functional hubs (Mueller et al., 2013) (which can be addressed by subject-specific parcellations (Dhillon et al., 2014)), inaccuracies in spatial alignment of the parcellation atlas (Smith et al., 2011; Allen et al., 2012), and the natural variability of cortical area size, which can vary by twofold or more across individuals (Amunts et al., 2000; Glasser et al., 2016; Bijsterbosch et al., 2018).

As one way to account for variability of functional locations in subjects, cluster-based parcellations have been proposed. In cluster-based parcellations, the parcellation is developed for the individual based on where activity is localized in the brain (Yao et al., 2015). Groupwise comparisons across personalized parcellations has its advantages and disadvantages; underlying networks are captured more effectively in personalized parcellations, but the fact that certain parts of the functional network are located in different anatomical areas may also be of interest, and this information is lost in personalized parcellations. Bijsterbosch et al. (2018) noted that differences in connectivity may indicate more about anatomical layout of functional regions than about differences in connectivity between those regions.

In order to address the shortcomings of hard parcellation techniques that fail to account for individual variations in functional hub locations and overlapping hubs, more advanced ICA parcellation techniques are able to parcellate overlapping areas (i.e., “soft” parcellations) using dual regression analysis or back projection (Calhoun et al., 2001; Filippini et al., 2009), which obtain subject-specific spatial maps using a group ICA maps (Bijsterbosch

et al., 2018). Seed correlation analysis is also used as a means of studying brain networks; however, this requires a prior assumptions about the location of the network, as one must predefine the areas to correlate.

Besides parcellations, another issue that can affect the quantification of brain networks is anatomical factors that confound the BOLD signal itself. Functional connectivity may increase as the result of non-neuronal fluctuations in the BOLD signal, such as blood pressure, head motion, or respiratory activity (Caballero-Gaudes and Reynolds, 2017). In most cases, the magnitude of changes from these confounding factors are greater than what one can expect from neuronal fluctuations in the BOLD signal, and so simple options such as wavelet despiking (Patel et al., 2014) or other bandpass filtering methods may be adopted. However, it is difficult to remove this global signal, and even with advanced preprocessing techniques (Kundu et al., 2012) and preventative efforts, trace effects of these non-BOLD artefacts usually remain.

1.4 Machine learning

1.4.1 Overview

"Machine learning" (ML) refers to a number of statistical models for identifying and generalizing patterns in data. Machine learning can either be unsupervised, which consists of methods of clustering unlabeled data (for instance, identifying friend groups in a social network, or communities in a brain network), or supervised, in which a model learns to associate patterns in the data with different labels (for instance, distinguishing between images of cats and dogs). This thesis focuses on supervised learning.

1.4.2 Training

In the typical supervised learning paradigm, one has a collection of data and a selected model. The data is then divided into two parts: training and test data. Most often, the majority of the data is used for training. The model is then "trained" on the training data for a number of iterations, often with datasets grouped into smaller batches. One update of the model's parameters over a single batch is called an "iteration", and when the model iterates over enough batches such that it has seen the entire training set, one "epoch" has

been completed. After the training algorithm has finished a set number of epochs, or has achieved 100% classification accuracy on the training set, it is then evaluated on the test set, and the final accuracy is reported.

One problem in machine learning is overfitting, in which the model overfits on the training set, such that it fails to generalize underlying patterns in the data. This is often the result of a model with too many parameters being trained on a training set that is too small, and thus the model is inappropriate for the data. It may also result from overtraining the model. To address this, data may be divided into three sets: training, test, and validation sets. The model does not train explicitly on the validation set, but it uses the validation set to measure accuracy at each epoch, stopping either when an accuracy threshold is reached, or saving a copy of the model on the iteration that produced the highest accuracy on the validation set.

Underfitting during training – in which the model fails to achieve 100% accuracy on the training set – is another problem, and is typically the result of too few parameters in a model, or patterns in a dataset too complex to effectively be characterized by the selected model (for instance, linear regression is likely insufficient to distinguish between pictures of cats and dogs).

Accuracy achieved on a particular division between training, test, and validation data is not deterministic, and it may change if that division changes. Furthermore, some machine learning models (especially neural networks) have stochastic elements that affect the final outcome of the model, even if the divisions remain the same. For this reason, a typical strategy in machine learning studies is to employ cross-validation (Kohavi, 1995), in which a number of independent models (usually a minimum of 10) are trained on different divisions of the data, with the divisions being deliberate such that each datapoint is included equally as often in the training, test, and validation sets. Another form of cross-validation, called leave-one-out testing, is typically employed for much smaller datasets; for a dataset of size n, n independent models are trained on a training set consisting of $n - 1$ datapoints, then evaluated on the one missing datapoint.

1.4.3 Overview of machine learning methods

Unsupervised machine learning methods refer to a number of techniques that may be applied to data without the need for labels, and these methods may be used for applications such as cluster identification (i.e., K-means clustering). Two popular unsupervised deep learning

models are Restricted Boltzmann Machines (RBMs) and variational autoencoders, which are used for data dimensionality reduction.

Support Vector Machines (SVMs) are supervised machine learning models capable of performing linear classification of high-dimensional data. SVMs find the best-fit linear division that separates data into two classes, possibly after artificially raising the dimensionality of this data. However, SVMs are only capable of finding linear divisions in data, usually failing to classify complex data such as photographic images.

Neural Networks refer to a number of models inspired by studies of neurons. Neural networks consist of layers of perceptrons, or units that receive a number of inputs and produce a single output. Deep learning is a subfield of machine learning that refers to the use of neural networks with different layers. Neural networks are particularly useful for characterizing complex and high-dimensional data, though extremely large amounts of data are necessary to train a neural network without overfitting. While neural networks have been the subject of research for decades, they were quickly popularized by the method of efficiently training neural networks (Hinton et al., 2006) and the subsequent display of their efficacy in the ImageNet competition (Krizhevsky et al., 2012). Since then, deep learning has been applied to many other fields of research, including MRI analysis.

In modern deep learning, two types of neural networks are very often used. Recurrent Neural Networks (Lipton, 2015) (RNNs) are particularly powerful deep learning models that use, as part of their input, the model output from previous data inputs. As such, they are best for learning sequences, such as semantic sentence interpretation (Karpathy and Fei-Fei, 2014). Convolutional Neural Networks (LeCun et al., 1999) (CNNs) are a powerful model for classifying images and videos, which encode the spatial organization of data by convolving adjacent pixels in successive layers, creating an abstract representation of objects in the process. This thesis employs convolutional neural networks extensively.

1.4.4 Convolutional neural networks to classify graphs

While the most popular application of CNNs has been in classifying 2D images (Krizhevsky et al., 2012), there has been a particular effort in recent years to adapt them to other kinds of data, such as 3D images (Maturana and Scherer, 2015), video (Karpathy and Fei-Fei, 2014), and audio-to-text conversion (Lipton, 2015). Because of their wide applicability in representing data such as proteins and social networks, much work has been done on clas-

sifying connected networks, including whole-graph classification, clustering, and node-wise classification (Bruna et al., 2014; Defferrard et al., 2016; Hamilton et al., 2017; Hechtlinger et al., 2017; Kipf and Welling, 2017; Nikolentzos et al., 2017). Graph classification is typically performed in one of three ways: graph kernels (Jie et al., 2013; Kriege et al., 2019; Nikolentzos et al., 2019), translating the graph to another representation (such as an image) before classifying it with a CNN (Tixier et al., 2017), or directly with convolutional neural networks (Kawahara et al., 2017).

Convolutional neural networks (CNNs) adapted for graphs have potent applications in the classification of brain connectomes. While other machine learning (ML) models have been developed for analyzing graph data (Jie et al., 2013; Kriege et al., 2019), they have often been designed to characterize general networks (such as social networks) rather than fixed-node matrix representations, and so are not ideal for brain connectomes. With its utilization of powerful deep learning structures (Kawahara et al., 2017; Brown et al., 2018), however, CNNs are among the most promising ML tools for the diagnosis and prognosis of neurological and mental health disorders using graph representations of the structure and function of the brain. This thesis largely focuses on the adaptation of CNNs to classify graphs.

1.4.5 Explainable AI and the black box problem

A common problem in deep learning is the "black box" problem. A black box is any system that receives an input and outputs a signal without knowledge of its internals, which are abstracted from users. Deep learning models, which often require millions or tens of millions of parameters, are abstracted from human understanding by their own complexity and are thus considered black boxes.

While the black box problem does not directly affect the performance of deep learning models, it creates difficulties when verifying whether a model is focusing on signal or confounding factors. In a well-known instance, Ribeiro et al. (2016) discussed a problem in which a classifier learned to reliably distinguish between pictures of wolves and huskies. However, a proposed explanation of the classifier revealed that, since wolves were mostly imaged with snow in the background, the classifier focused on snow rather than the face of the wolf. They concluded that, in spite of high accuracy, this would not be an appropriate classifier to rely on in a real-world setting.

In machine learning for scientific discovery, interpretability is arguably just as important as

classification accuracy. The need for explainable machine learning models in a clinical setting has previously been discussed (Gottesman et al., 2019). Clinicians need to fully understand the decision-making process of an automated diagnosis if they are to eventually rely on it. AI models that make a linear, understandable decision-making process are called "expert systems". These often rely on human-readable information, such as the diagnostic history of an electronic health record. However, such systems would not be capable of making use of more complex datasets that are not always human-readable, such as medical images or genetic records.

Deep learning models have been shown to be capable, at least to a degree, of making sense of complex datasets, in a way that an explainable expert system (Gottesman et al., 2019) would not, in applications like whole-brain MRI diagnostics (Kawahara et al., 2017; Khosla et al., 2018; Leming and Suckling, 2020a,b)). Unlike expert systems, deep learning models' decision-making processes are too complex for human understanding (i.e., the black box problem). Because of the need for clinicians to explain their decisions, this would make deep learning models of limited value. There has been great effort in visualizing deep learning models in other contexts in the hope of making them explainable. These methods include occlusion, gradient class activation mapping, and activation maximization (Zeiler and Fergus, 2013). While these methods fail to reveal the exact decision-making process used to make classifications, they are capable of showing which parts of the input data are taken into account for the classification. Use of such techniques can make deep learning models more explainable, and thus more useful in an eventual clinical context. But while such methods help explain machine learning models, the full extent of these techniques, and the exact interpretation of any visualization techniques in a scientific context, is still the subject of ongoing research.

1.4.6 Methods of visualizing deep neural networks

The black box problem has motivated a sub-field of research into methods of analyzing and visualizing deep learning models. Most of these techniques were originally developed for 2D image recognition and adapted later to abstract data types. The interest in visualization methods has motivated different sub-fields of deep learning, such as generative models (Goodfellow et al., 2014), unsupervised object localization (Zhou et al., 2015b; Oquab et al., 2015; Cinbis et al., 2015; Pinheiro and Collobert, 2015; Bergamo et al., 2014; Oquab et al., 2014), and deconvolutional neural networks (Zeiler et al., 2010).

Of particular interest in this thesis are those methods that show which parts of input data contribute to classification accuracy – i.e., salience – with the aim being that this elucidates group differences, as well as methods that analyze how the machine learning model itself sorts large datasets during classification, in order to quantify whether a classifier focuses on signal or confounding factors. Three general methods are discussed here: class activation maps, occlusion, and activation maximization.

Class activation maps

Class activation maps (CAMs) measure the influence of localized areas of input on the final accuracy calculation (i.e., "salience"). In the context of photographic image recognition, this is usually interpreted as a measurement of human fixation; for instance, in a classifier that distinguishes between pictures of cats and dogs, the CAM should highlight the shape of the cat or dog in the input image. In its earliest iteration (Simonyan et al., 2014), CAMs were estimated as the derivative of the deep learning function with respect to the input image, approximated as a first-order Taylor series. This, in effect, showed the degree to which each part of the input image altered the final classification. However, the first implementations of CAM estimation offered noisier results before algorithmic improvements were developed. Zhou et al. (2015a) developed CAM estimation further with an algorithm that was more effective at object localization but was only used for specialized CNNs with no fully-connected layers. The later innovation of Selvaraju et al. (2017) generalized this to Guided Grad-CAM, a version of this class activation mapping algorithm that could be applied to a wider variety of deep learning models.

Developments in CAM estimation have re-formed it into an object segmentation tool (Zhou et al., 2015a). Many later developments (Li and Yu, 2018) expanded on it by employing contour-based methods for object segmentation. However, when applicable deep learning models are adapted to other, abstract data types, such object-segmentation-focused salience detection may not be ideal. The parts of input data that affect the output the most would likely not be able to be clustered together in the way objects in 2D images are, as other data may be more abstract than 2D objects. This thesis makes extensive use of CAM algorithms, but because of the irrelevance of object segmentation to brain connectomes, I opt for an earlier implementation (Selvaraju et al., 2017), Guided Grad-CAM, rather than the current state-of-the-art.

Guided Grad-CAM obtains the class-discriminated gradient (i.e., first-order derivative with

respect to a class) of a neural network with respect to the feature maps of a convolutional layer. The gradients are then average-pooled. This represents the salience of particular feature maps for a particular class. A weighted combination of all of these forward activation maps are then added and followed by a rectified linear unit (ReLU, i.e. the absolute value), to obtain a coarse heat map of the same size as the convolutional feature maps (Selvaraju et al., 2017).

Occlusion

Occlusion (Zeiler, 2012) consists of occluding local areas of data and testing which of these lowers classification accuracy the most, effectively showing which areas of the image are most important in this classification. Like CAMs, it is a salience detection technique, and it is advantageous in that it more directly tests for salience. Unlike CAMs, occlusion does not actually involve direct analysis of deep learning parameters at all, but instead works by editing input data; this makes it an applicable method for analyzing black box models in general.

Occlusion can have several variations that have been applied creatively to deep learning: Zhou et al. (2015b) proposed an interesting variation on occlusion, in which the the pixel of an input image that caused the least accuracy was erased until the image was inaccurately classified, and Bergamo et al. (2014) used it as a means of unsupervised object segmentation. However, occlusion is also more computationally intensive than CAM estimation, somewhat limiting its use. Furthermore, the elimination of certain parts of input data may cause the deep learning model to act unpredictably, as it is known that random noise as inputs may output unpredictable results in deep learning models (Szegedy et al., 2014; Goodfellow et al., 2015).

Activation maximization

Activation maximization (Erhan et al., 2009) consists of recording which input data maximize which units in a particular hidden layer of a deep learning model. Typically, one finds the activation maximization of convolutional layers rather than dense layers, as convolutional layers maintain a level of stratification that helps with analysis; for instance, activation maximization of convolutional layers in 2D image recognition can render visuals of abstract shapes and textures of particular input objects within the network. Activation maximization

is used to assess how data is organized by the model; for instance, certain subclasses of data are often found to maximally activate different filters in convolutional layers.

1.4.7 Machine learning in medical imaging of the brain

Machine learning has been applied to medical imaging since the 1990s (Zhang et al., 1994), though in more recent years, the industrial deep learning boom has substantially affected medical imaging; the number of publications about machine learning in medical imaging has increased substantially since 2012, with CNNs being the most published about by far (Litjens et al., 2017; Shen et al., 2017a).

Machine learning has been applied to medical imaging for segmentation (Perone and Cohen-Adad, 2019), detection (Tajbakhsh et al., 2015; Tajbakhsh and Liang, 2015; Shin et al., 2016), object classification (Singh and Singh, 2017), and single-subject phenotypic classification (Arbabshirani et al., 2017). Such applications may generally be divided into two categories: ones for which human experts can achieve near-perfect accuracy, and ones which cannot due to the non-discovery of consistent biomarkers. The first category tends to see applications related to image segmentation, such as skull-stripping and tumor segmentation, and even extends into cancer diagnoses (Munir et al., 2019) by analyzing images of tumors or cells; because expert human interactors can achieve near-perfect accuracy with such images, it is established that the necessary information to succeed at the classification task is present in the given dataset. In the second category are image-based diagnostics, often of degenerative or mental disorders, for which human interactors cannot readily make a successful classification given this data, since the given data is usually not the primary source of such a diagnosis in the first place. For instance, clinicians and radiologists would not diagnose autism based on brain images, but rather by behavioral markers; thus, it is unknown whether a biomarker exists for autism (Plitt et al., 2015). The present thesis largely focuses on single-subject phenotypic classification using MRIs, which is in the latter category.

Several different classes of machine learning models are in popular use in medical imaging; however, because of their applicability to images, CNNs are a popular choice. According to Litjens et al. (2017), "out of the 47 papers published on [whole-image] classification in 2015, 2016, and 2017, 36 are using CNNs, 5 are based on [autoencoders] and 6 on RBMs". A review of more years in Arbabshirani et al. (2017) also revealed a widespread use of SVMs, which are more applicable to smaller datasets and suffer less from the black box problem, though SVMs require specialized feature extraction methods, which are often study-specific,

and are often unable to characterize nonlinear patterns in data.

Machine learning studies in brain connectivity have been used for single-subject classification of bipolar disorder, attention deficit hyperactivity disorder (ADHD), mild cognitive impairment (MCI), schizophrenia, autism, attention deficit hyperactivity disorder (ADHD), and Alzheimer's disease (AD) (Du et al., 2018), with studies variously using functional and structural data for classification, depending on the task. Certain disorders, such as autism and ADHD, are mainly characterized by behavioral symptoms, with structural and functional brain differences still being an active area of research. However, even these behavioral symptoms are difficult to characterize; individuals with autism are not characterized by the same profile, and their symptoms are known to change over time (Lord et al., 2000). The causes of ADHD are still unclear and are likely a number of possible factors, including heredity, brain chemistry, and malnutrition (Biederman, 2005; Dey et al., 2014). Naturally, an incomplete understanding of the disorders themselves does not help with the technical difficulties of classifying images of the brain.

Most of these studies rely on small training datasets. As noted by Arbabshirani et al. (2017) and Katuwal et al. (2015), machine learning models for whole-brain MRI classification generally perform better on small, single-site datasets than on large, mixed-site datasets. This would seem to contradict the conventional wisdom in machine learning that larger training datasets aid model performance. A thorough explanation of this phenomenon has not been offered, but it is typically assumed that mixed-site datasets add confounding factors and variations that are difficult for a model to characterize.

The shortage of health data to aid in machine learning algorithms has led to several efforts in industry at amassing data already present in health records, such as Google's Project Nightingale (Copeland, 2019) or IBM Watson Health's partnerships with hospitals (Quach, 2018), though such efforts are routinely plagued by controversy, either due to concerns with diagnostic accuracy (IBM Watson) or data privacy (Project Nightingale). The collection of larger datasets in the research world, such as UK BioBank, holds more promise for studies of big data in the near future.

1.5 Big data in medical imaging

1.5.1 Overview

Like many biological applications, machine learning studies in MRI are limited by sample size, as MRIs are comparatively expensive and labor-intensive to acquire. With growing interest in big data in MRI (Smith and Nichols, 2018), wider initiatives to acquire large datasets has allowed for training sets in the hundreds (Abraham et al., 2016) or thousands (He et al., 2018), though even these are orders of magnitude smaller than those used in mainstream computer vision (Wu et al., 2019).

1.5.2 Role in machine learning

There have been methods of overcoming the limitation of small sample sizes in medical imaging, such as data augmentation (Hussain et al., 2017), the use of leave-one-out classifiers (Anderson et al., 2011b; Jang et al., 2017), the use of simulated data (Meszlényi et al., 2017), or the development of methods that specialize in training on lower sample sizes (Liu et al., 2014; Akkus et al., 2017; Gibson et al., 2018; Han et al., 2017; Shen et al., 2017a). However, Arbabshirani et al. (2017) observed that most results that showed extremely high accuracy ($> 90\%$) were most often performed on samples of less than 100. This could be due to overfitting of newly developed models, homogeneity of samples, the use of leave-one-out classification (which, though it may seem appropriate for small datasets, can prove statistically unsound (Kohavi, 1995)), or unknown factors. In any case, this brings into question the generalizability of many such models, and it is likely another expression of the statistical problems associated with low sample sizes (also called "power failure") in neuroscience (Button et al., 2013; Nord et al., 2017).

This necessitates the use of big datasets. In recent years, this has become a more viable option, as there have been several larger initiatives – for instance, the UK Biobank and ABCD – that have collected thousands of datasets from different facilities that centrally work to minimize site differences and make the data as high-quality and homogeneous as possible. The UK Biobank is especially notable for housing extremely detailed metadata about its subjects, allowing for many big-data psychological studies to be performed on the basis of this metadata alone. Such large databases, however, can be lacking in variety of patients with a clinical diagnosis that is often of particular interest to researchers, though

there are several medium-sized ($100 < N < 5000$) databases that are dedicated to the study of particular disorders (i.e., ADNI for Alzheimer's and ABIDE for autism).

Large imaging databases are often composed of smaller batches collected by individual research groups, which are then amalgamated in repositories, such as NDAR and OpenfMRI. These small MRI datasets, however, often have variations that make it extremely difficult to compare subjects in cross-dataset studies. Such differences are not only limited to MRI hardware, parameter, and population differences, but subtle variations in clinical and diagnostic practices depending on the disease being studied, and preprocessing practices of the holding database. Furthermore, standard practices in brain imaging often calls for a level of human interaction, if only to perform quality control, when preprocessing data, and for collections of this size such standard preprocessing practice would not be viable.

Nonetheless, noisy repositories, in aggregate, house much data of interest, and though the many factors noted above may make it impractical to compare them using conventional methods, deep learning was designed to classify such data with high levels of noise and variation. Deep learning does not resolve all such concerns – for instance, one has to be sure that high accuracy is not simply due to gross imbalances in classes or site differences – but, in theory, it does lessen the need to explicitly model and regress artifacts such as motion or signal weakness, as long as the model is unable to utilize such factors to improve classification accuracy.

1.6 Connectivity in different cohorts

Brown and Hamarneh (2016) provides an overview of previous efforts in brain connectome classification on different phenotypic groups. A direct comparison between deep learning studies is complicated by differences in machine learning models; differences in datasets; preprocessing practices; division between training, test, and validation sets within the same dataset; and differences in metrics used to validate one's machine learning model. In this section, I summarize previous efforts to classify based on phenotypes considered in the present work (sex, task, autism, and depression); when such studies are sparse or nonexistent, I discuss factors that would likely affect such studies.

1.6.1 Task fMRI

The functional differences found between task-based and resting-state fMRI may be among the most consistent occurrences in fMRI studies. Corbetta and Shulman (2002) first discovered the dorsal and ventral attention networks (Vossel et al., 2014), which were respectively concerned with voluntary focus on features and switches in attention or unexpected stimuli. As noted by Fox et al. (2005), when performing simple memory tasks in a fMRI, the response commonly observed is increased activity in certain frontal and parietal cortical regions (Cabeza and Nyberg, 2000; Corbetta and Shulman, 2002) and decreased activity in the posterior cingulate, medial and lateral parietal, and medial prefrontal cortex (Gusnard et al., 2001; Simpson et al., 2001; Shulman et al., 1997; McKiernan et al., 2003; Mazoyer et al., 2001), which form the default mode network; the intensity of this response was proportional to the intensity of the task. Fox et al. (2005) identified two widely distributed, anticorrelated networks in the brain that exist in the resting state but intensify during tasks.

The default mode network has been consistently identified as a marker of resting-state connectomes since it was first described in Raichle et al. (2001), and other brain networks, including some emblematic of particular tasks, have been identified as well (Smith et al., 2009). Using fMRI and diffusion-weighted MRI, Yoldemir et al. (2015) distinguished, with 79% accuracy, between seven functional tasks using the fMRI timeseries. On the whole, classification of resting-state and task-based fMRI is underexplored, though Zhang et al. (2016a) recently used sparse representations to distinguish between task- and resting-state fMRI in the Human Connectome Project data, achieving 100% accuracy and distinguishing between subjects by identifying the presence of the default mode network, though this was done on a dataset consisting of only 60 subjects (even though the data collected on each subject was robust and detailed). There has also been work in using deep learning to decode different brain states in individuals (Koyamada et al., 2015), and Li and Fan (2018) creatively applied recurrent neural networks to decode brain states in single-subject fMRI as they changed over the course of a timeseries.

1.6.2 Sex

Evaluation of male-female brain differences are of general and widespread interest in neuroscience, but this has uniquely complicated the subject: the frequency with which it is studied, combined with the statistically unsound use of small sample sizes (Button et al.,

2013; Nord et al., 2017) in studies that use different analytical methods, has created a very inconsistent picture across the board with regards to brain functional and structural differences between sexes; another view of this, however, is simply that male-female differences are highly complicated and the literature reflects that (Ruigrok et al., 2014).

Functional differences between the sexes are debated and findings generally vary, depending on which aspects of it are studied. A review by Sacher et al. (2013a) cited evidence for functional sex differences in emotional and visuospatial processing, but noted that, up to that point, it was rarely considered as a covariate in fMRI studies, and thus more rarely in resting-state fMRI. Noted differences in emotional and visuospatial processing include arousal differences in the bilateral amygdala and hypothalamus (Hamann et al., 2004; Takahashi et al., 2006; Mackiewicz et al., 2006), the right cerebellum, and the posterior and superior temporal sulcus (Takahashi et al., 2006), as well as hemispheric differences, in response to various emotional stimuli. Men and women also differed in right hemisphere activation in response to visuospatial tests (Gur et al., 2000), and differing activations in the superior parietal lobule and the inferior frontal cortex in response to mental rotation tasks (Hugdahl et al., 2006). In a task in which subjects were presented with emotional faces, men showed higher activation in limbic and prefrontal regions and women higher activation in the right subcallosal gyrus (Fusar-Poli et al., 2009). In response to angry and fearful faces, men show higher activation than women in the visual cortex and the anterior cingulate gyrus (Fischer et al., 2004).

Later studies were performed on large samples of resting-state data rather than smaller samples of task fMRI with different emotional stimuli, mitigating the concerns raised in Button et al. (2013) and Nord et al. (2017). Sex differences have been analyzed in large-sample studies of children and adolescents (Gur and Gur, 2016; Gennatas et al., 2017; Wierenga et al., 2017) and the release of large samples of data in the UK Biobank (Ritchie et al., 2018) (2750 females and 2466 males). (Because this work uses both resting-state and task fMRI, it is reasonable to include both of these in my analysis and hypothesis formation.) High-sample-size studies of resting-state fMRI differences between sexes have not typically implicated differing activations in very localized portions of the brain; generally, however, studies have reported, through different measurements, higher local functional connectivity in women than in men (Tomasi and Volkow, 2011b; Gur and Gur, 2016), though different analytical methods can also make these studies particularly tricky to compare, since individual studies report arbitrary graph-theoretical measurements or different network ROIs in seed-based analysis.

Unlike function, differences in brain structure between females and males has been extensively documented (Ruigrok et al., 2014), both in the developing brain and later in life. Earlier, smaller-scale studies reported that, in the beginning of life, one-year-old boys have bigger brains by about 10 percent, but girls have bigger structures in the brain stems (Giedd et al., 1997) (50 males and 70 females, aged 3–18), and in both sexes the right hemisphere is generally bigger than the left (Baibakov and Fedorov, 2010) (30 males and 30 females, age 12). The sizes of structures are not proportionally bigger across age and sex (allometry) (Giedd et al., 2012). Development of certain structures, such as the amygdala and hippocampus, are different in early adolescent years for females and males (Uematsu et al., 2012) (58 males, 53 females, studied from 1 month to 25 years).

Among the most notable findings of large-scale studies of the developing brain, which represent samples of total size 1929 males and 2065 females aged 3 to 23 years (of whom 745 males and 826 females were analyzed functionally (Gur and Gur, 2016)) were the following: (1) a steeper increase in white matter volume in males than in females during puberty, especially in the frontal lobe (Lenroot and Giedd, 2006; Gur and Gur, 2016); (2) bilaterally larger hippocampal volume in females after puberty, correlated with memory tests, and equal volumes before (this effect was not seen in the amygdala, however) (Gur and Gur, 2016); (3) in their resting-state functional networks, when they were separated into modules, males showed higher between-module connectivity and females showed higher within-module connectivity (Gur and Gur, 2016); use of SVMs to classify males and females achieved 63% accuracy using only their cognitive profile but 71% accuracy using functional connectivity data (Gur and Gur, 2016), improving on a previous accuracy of 65% accuracy for a similar age group based on functional data from the human connectome project (Casanova et al., 2012) (74 females and 74 males, age 21); (4) the functional connectome shows greater modularity in females and the structural connectome has greater modularity in males (Gur and Gur, 2016); (5) Throughout the brain, females have lower gray matter volume but higher gray matter density than males (Gennatas et al., 2017); (6) with regards to the volume of several key brain structures, including cerebral white matter and cortex, hippocampus, pallidum, putamen, and cerebellar cortex, males showed significantly greater variance than females (Wierenga et al., 2017). Many of these results were supported by an earlier study (Tomasi and Volkow, 2011a) that used a notably large sample size of young adults (336 females and 225 males, aged 18–30 years), finding that women have higher local functional connectivity density and higher gray matter density than men.

The findings of Ritchie et al. (2018), which represent a much older sample of 2750 females and

2466 males, aged 44 to 77 with a mean of 61.7, found that males have higher brain volumes, surface areas, and white matter fractional anisotrophy, whereas females have higher cortical thickness and white matter complexity. With regards to resting-state functional connectivity, females had stronger connectivity in the default mode network and stronger connectivity for males in the sensorimotor cortices; this may be another expression of the findings of Gur and Gur (2016) and Tomasi and Volkow (2011b) with regards to females having higher within-network/local connectivity (considering the default mode network is one of the most prominent networks in the brain), but given that the results are presented differently, it remains difficult to tell. Supporting the findings of a younger cohort in Wierenga et al. (2017), males also had greater variation in these measurements. With regards to all measurements taken, however, there is considerable overlap between groups.

Studies of sex differences in brain structure and function are underpinned by a wide body of literature concerning cognitive and emotional differences between males and females that may coincide with these functional and structural differences. Studies have shown that males outperform females on spatial and motor cognitive tasks, while females outperformed males on nonverbal reasoning and emotional identification (Gur and Gur, 2016). Males are generally more physically aggressive (Archer, 2004) and more interested in things rather than people (Su et al., 2009), while females more often display neuroticism (Schmitt et al., 2008) and agreeableness (Costa et al., 2001) and are more interested in people rather than things (Su et al., 2009). With regards to neurological and psychological illness, females show a higher prevalence for Alzheimer's (Mazure and Swendsen, 2016) and major depressive disorder (Rutter et al., 2003; Gobinath et al., 2017), while males show higher prevalence for autism (Baron-Cohen et al., 2011), schizophrenia (Aleman et al., 2003), Tourette syndrome (Bitsko et al., 2014), and dyslexia (Arnett et al., 2017). Cognitive-functional studies have found differing functional responses in men and women in response to menstrual cycles and emotional stimuli (Stevens and Hamann, 2012; Sacher et al., 2013b). Past ML studies using methods ranging from support vector machines to CNNs have achieved sex classification accuracies between 65% and 87% (Casanova et al., 2012; Satterthwaite et al., 2015; Gur and Gur, 2016; Zhang et al., 2018), depending on the dataset and methods used.

While there are evidently differences in many aspects of the male and female brain, nearly all of these studies note the distributional overlap and differences in variation between groups. This indicates that any machine learning tests on the brain will (1) likely never reach perfect accuracy, especially if they include a younger age group, and that any study approaching perfect accuracy should be approached with a degree of skepticism; (2) be strengthened with

additional information about brain structure, function, and cognitive tests; and (3) probably never find one localized biomarker of male-female differences by analyzing only MRI brain structure and function, though this may be changed with future improvements in spatial and temporal resolution of MRI.

1.6.3 Autism

Although there have been multiple reports of structural brain differences in autism (Redcay and Courchesne, 2005; Stanfield et al., 2008; Nickl-Jockschat et al., 2012a), this has not been substantiated by a wider-scale analysis (Haar et al., 2016). Recent longitudinal analyses in brain volume growth have shown that age trajectories in autism development had high inter-individual variability (Ha et al., 2015; Lange et al., 2015; Wolff et al., 2018), and this complex age/disease interaction renders autism even more difficult to study.

However, autism has been consistently associated with differences in brain function (Müller et al., 2008; Simas et al., 2015a). Efforts to find differences in functional connectivity between autistic patients and control groups have characterized autism as a disorder exhibiting under-connectivity and thus greater segregation of functional areas (Just et al., 2004; Cherkassky et al., 2006; Kennedy and Courchesne, 2008; Assaf et al., 2010; Jones et al., 2010; Weng et al., 2010). Other studies, mostly of children and adolescents, found evidence of over-connectivity in specific areas of the brains of subject with autism (Cerliani et al., 2015; Chien et al., 2015; Delmonte et al., 2013; Di Martino et al., 2011; Nebel et al., 2014a,b), finding hyperconnectivity in the posterior right temporo-parietal junction (Chien et al., 2015) and in striatal areas and the pons (Di Martino et al., 2011; Delmonte et al., 2013). A recent review (Hull et al., 2017) posited that autism is likely characterized by a mix of these traits.

Autism has been characterized by several cognitive theories (Lai et al., 2014); the three dominant ones still actively researched (Hull et al., 2017) are the Weak Central Coherence theory (Frith, 1989, 1996; Happé and Frith, 2006), the Executive Dysfunction hypothesis (Levy, 2007; Romero-Munguía, 2013; Fishman et al., 2014), and Theory of Mind (Baron-Cohen, 1988c,a,b; Baron-Cohen et al., 1994; Baron-Cohen, 2004). In early infancy, however, autism is particularly difficult to diagnose, since most diagnostics rely on behavioral measures (Shen and Piven, 2017), though there has been success in identifying behavioral markers in a particularly affected subgroup of infants in the 6–9 month period, such as unusual visual fixations and lack of intentional communicative acts (Bryson et al., 2007; Rogers et al., 2014).

Previous efforts to classify functional connectivity in autism on small datasets have achieved accuracies that have been described as "modest to conservatively good" (Hull et al., 2017), though such methods have had trouble replicating on different data (Jung et al., 2014; Price et al., 2014; Iidaka, 2015). Generally, studies achieved classification accuracies that widely varied depending on the modality used, sample size, data quality, selected methods, and diagnostic criteria. More recently, the application of convolutional neural networks to the substantially larger ABIDE I and II datasets have achieved achieved 68% to 77.3% classification accuracy (Subbaraju et al., 2017; Brown et al., 2018; Heinsfeld et al., 2018; Khosla et al., 2018). A recent study (Hazlett et al., 2017) of 106 high-risk infants between 6-12 months linked brain volume overgrowth to the emergence and severity of autism symptoms, using a deep learning algorithm capable of predicting autism with 81% specificity and 88% sensitivity using brain surface information. Another study by the same group (Emerson et al., 2017) found that autism could be predicted in 59 6-month-old infants with 81.8% sensitivity using functional imaging. In the general population, efforts in single-participant classification of autism from MRI data have had mixed results (Anderson et al., 2011a; Barttfeld et al., 2012; Nielsen et al., 2013b; Jung et al., 2014; Iidaka, 2015; Plitt et al., 2015; Tejwani et al., 2017), with studies rarely exceeding 80% classification accuracy (Hull et al., 2017). Again, however, this varies substantially by modality and which site data were collected (Katuwal et al., 2015). Nielsen et al. (2013b) found that multisite functional connectivity classification in autism outperformed chance, but the highest accuracy obtained was 60%. It also found that accuracy was significantly higher for sites with longer BOLD imaging times; this is compared with a single-site study (Anderson et al., 2011b) that saw around 80% accuracy for whole-brain autism classification and 91% for subjects under the age of 20, though this used a sample size of 80, and, as mentioned earlier, such disparities are common between high-sample-size and low-sample-size machine learning studies (Arbabshirani et al., 2017). In a recent study, Eill et al. (2019) performed a classification on individuals with autism and neurotypical controls using structural MRI, DWI, and fMRI data, finding that features derived from fMRI provided the highest accuracies with an SVM classifier. They did, however, encounter the issue of fMRI feature extraction simply producing more variables than its structural counterparts, offering the machine learning model more information to work with, although attempts were made to mitigate this issue.

1.6.4 Depression

MDD has been studied extensively (Zhang et al., 2011a; Bora et al., 2013; Graham et al., 2013; Li et al., 2013; Roiser and Sahakian, 2013; Singh and Gotlib, 2014; Qiu et al., 2015). Using different methodologies, different studies and meta-analyses have implicated case-control differences (both in terms of structure and function) in many different parts of the brain (Kaiser et al., 2015; Mulders et al., 2015), and others have shown only limited areas of difference (Bora et al., 2013). There are several possible explanations for this. The first is that MDD is a complex disorder and each methodology uniquely captures a different aspect of the disorder. The second is that many methods used potentially capture spurious differences in the data. The third is that MDD is a system-wide disorder and different methods implicate specific parts of the brain, each partially illuminating a deeper, more widespread effect. Another explanation for the dissimilarities is the slight differences in the datasets studied.

On small samples, there has been marked success in classifying major depressive disorder, with accuracies up to 100% (Zeng et al., 2012; Rosa et al., 2015; Sato et al., 2015; Ramasubbu et al., 2016; Wang et al., 2017; Yoshida et al., 2017), and overall accuracies of 77% in subclinical depression (Modinos et al., 2013). While there were found to be average differences in subcortical volume and white matter integrity in UK Biobank participants (Shen et al., 2017b), there have been no published efforts to classify subclinical depression on the BioBank's large functional MRI datasets.

1.7 Research Aims

The overarching purpose of deep learning for MRI classification is twofold. First, there is a scientific interest: if a deep learning model is able to classify a MRI data by a certain mental disorder, then studying this deep learning model would undoubtedly aid in understanding of this disorder. The second interest is in future clinical application: given high enough classification accuracy, this research has a clear potential application in automated clinical diagnosis and prognosis.

The aim of the research presented in this thesis is also twofold: first, to leverage big MRI data to classify phenotypes with as high a performance as possible. This involves obtaining such data and designing a deep learning scheme that can classify it as well as possible. Unique to

science, however, this also means ensuring that such classification is not due to confounding factors (such as, with MRI, head motion or intracranial volume), but rather to signals such as fluctuations in the BOLD signal. The second aim is to study these models to aid understanding of the studied phenotypes. This involves adapting deep learning visualization methods used to detect which parts of input data drive classification, thereby shedding light on the nature of phenotypic differences.

1.8 Thesis outline

The research of this thesis revolves around the use of deep learning for the phenotypic classification of whole-brain MRIs. Three major deep learning studies are presented throughout the thesis (in Chapters 4, 7, and 8), with other chapters outlining supporting material or specific methods to improve and elucidate these results.

Chapter 2 details my efforts in amassing and preprocessing an extremely large dataset from multiple different databases, as well as the practical challenges, otherwise out-of-place in the scientific narrative, in working with this data and designing deep learning models.

Chapter 3 presents an initial foray into connectivity analysis by presenting two distinct methods of analyzing connectomes, though neither rely on machine learning. The first of these, an analysis of "normative pathways", is a self-contained study on a smaller dataset that shows an interesting application of graph theory to functional connectomes, which is applied to adolescents with MDD; Chapter 9 later discusses means of applying such methods in future deep learning models. The second of these methods is a structural connectivity metric estimated from T1-weighted MRIs, which is later used in Chapter 8 for a deep learning study.

Following this, Chapter 4 presents a deep learning study on the massive aggregated dataset, using a convolutional neural network to classify functional connectomes by autism, sex, and resting state/task. This tested the viability of deep learning on such data. The study used ensemble CNNs in a cross-validation scheme, using an original deep learning encoding method that was partially inspired by an earlier framework called BrainNetCNN (Kawahara et al., 2017); it also took advantage of the depth of CNNs to encode multi-band wavelet correlation functional connectomes. This is accompanied by an analysis of the ensemble models and classification accuracy across different tasks and collections.

Subsequent chapters focus on analysis and improvement of the deep learning scheme from Chapter 4; however, to focus research efforts, deep learning studies presented after this point focus on only one of the classification tasks. Chapter 5 presents a method of analyzing the clustering of data throughout the deep learning ensemble models, as a means of measuring the degree to which it focused on confounding variables. Chapter 6 presents a method of mitigating this problem with an improved dataset balancing scheme. Chapter 7 focuses on addressing the "black box" problem by presenting two methods of analyzing the salience of input edges and networks, honing in on one classification problem in doing so: sex, specifically in UK BioBank data (a large subset of the full dataset collected). Finally, Chapter 8 homes in on another classification task, autism versus controls, by encoding both functional connectivity data and the structural connectivity metric presented in Chapter 3, and presents methods of using graph theoretical metrics for more advanced analysis of edge salience.

Chapter 9 ends the thesis with a discussion of the implications of this research; the limitations of big data, machine learning, and advanced statistical methods in whole-brain MRI classification; caveats in interpreting visualization metrics; failed research directions that I attempted; and potential future directions of this research. Purely neuroscientific implications of this research are relevant to the context of individual chapters, so this discussion mainly focuses on methodology, which is the overarching contribution of this thesis.

1.9 Motivation and Themes

In general, the boom of research in computer science has led to the creation and rapid development of many novel analysis methods that have been adopted by other scientific fields. However, because deep expertise rarely spans across two fields, these methods are often misapplied, either by pure computer scientists or statisticians failing to respect the complexities of another field, or by field experts using an off-the-shelf tool created by computer scientists.

This tendency is especially true of medical image analysis. For instance, a major pitfall discussed in this work is the emphasis on high test set accuracy in machine learning studies. In photographic image classification or speech recognition, high test accuracy is often the central goal, and so medical imaging researchers in whole-brain MRI or EEG classification often equate high accuracy with superior performance. However, such accuracy may be due to confounding factors such as motion, head size, or age, and because such factors do not

play a part in photographic images, they are often not fully considered in other fields.

Thus, a significant underlying theme of this study is the proper adaptation of computer science methods to the analysis of MRI brain data, respecting the complexities of both fields while exploiting unique properties of computer science methods to reveal novel neuroscientific insights. To this end, this thesis prototypes the use of "big data" in medical image analysis, collecting large samples of MRI data and adapting other metrics to analyze this data, ranging from deep learning visualization methods to pure graph theory, while carefully considering these methodologies in the context in medical image analysis.

The choice of which neuroscientific questions to address in this thesis have largely been driven by available data. Even with a sufficiently large dataset, there are limited means of applying it to supervised learning. Supervised learning requires data to be labeled. However, in collecting this data, I found large datasets to be inconsistently labeled; and, often, in the case of large datasets that do have consistent labels (such as the UK BioBank), the labels may not be of particular scientific interest. Furthermore, labels that are of scientific interest are not guaranteed to produce meaningful results. For instance, I do not present classification results based on factors such as age, which had too many confounding factors across data, and hallucination and subclinical depression data, which was present in the UK BioBank but failed to yield a meaningful classification accuracy when tested for. The only three interesting labels found consistently across a sufficiently large dataset, which also produced significant classification accuracy, were sex, autism, and resting-state/task. Thus, the classification tasks presented focus on those three labels.

Many methods presented in this thesis are targeted to datasets of very large sample sizes, and so would have relatively limited application to many studies in medical imaging today. One instance of this inapplicability is the heavy use of balancing algorithms, presented in Chapter 6, as a method of regressing out confounding factors. In some cases, this required discarding 90% or more datasets in model training; doing so on much smaller sample sizes would likely lose any statistical power, let alone their application to deep learning models, which require large amounts of data to be effective. However, with many ongoing big data initiatives in MRI, such methods will undoubtedly become more and more useful in the future.

Chapter 2

Amassing and processing large datasets

Throughout this thesis, I focus on the application of deep learning for whole-brain phenotypic classification to a large, mixed-site MRI dataset. This chapter provides an explanation of the acquisition and preprocessing of this data, practical details of the implementation of the deep learning framework, and memory management and computing challenges inherent in this study. Details in this chapter are applicable to all following it, except when otherwise noted.

I first present dataset collections, acquisition techniques, and labeling. I then detail general techniques for preprocessing fMRI data (note that structural data processing is the subject of Chapter 3). I then outline the implementation behind the deep learning framework used to classify this data, including the specific software libraries used. Aspects of these are expounded upon in later chapters as well, though the emphasis in those contexts is on design and scientific applications, while this chapter details implementation and practical challenges that do not otherwise fit in a scientific narrative.

2.1 Data acquisition

A very large number of structural and functional MRI datasets were collected from nine different sources: OpenFMRI (Poldrack et al., 2013; Poldrack and Gorgolewski, 2017); the

Alzheimer's Disease Neuroimaging Initiative (ADNI); ABIDE (Di Martino et al., 2014); ABIDE II (Di Martino et al., 2017); the Adolescent Brain Cognitive Development (ABCD) Study (Casey and Dale, 2018); the NIMH Data Archive, including the Research Domain Criteria Database (RDoCdb), the National Database for Clinical Trials (NDCT), and the National Database for Autism Research (NDAR) (Hall et al., 2012) (note: due to the difficultly in distinguishing between these three datasets when downloading them, I refer to these collectively as "NDAR"; while ABCD was downloaded from the NIMH database as well, this was significantly easier to distinguish); the 1000 Functional Connectomes Project (Dolgin, 2010); the International Consortium for Brain Mapping database (ICBM); and the UK Biobank; I refer to each of these nine sets as *collections*. These collections contain both resting-state and task-based fMRI from both control and various types of test subjects. OpenFMRI, NDAR, ICBM, and the 1000 Functional Connectomes Project are each collections that comprise different datasets submitted from unrelated research groups; ADNI, ABIDE, ABIDE II, ABCD, and the UK Biobank are collections that were acquired as part of a larger research initiative, and, while many of them did collect data from different sites, there was deliberate effort to minimize site differences. The numbers of subjects, total numbers of functional datasets, and connectomes derived from each, as well as phenotypic distributions, are shown in Table 2.1.

2.1.1 Dataset descriptions and labeling

Following are in-depth descriptions of each dataset and the means used to gather labels from each one. The labels most commonly occurring across each were (1) sex; (2) resting-state/task; and (3) age. Several datasets also contained large numbers of labeled data for autism.

NIMH (including NDAR and ABCD) The NDAR datasets include 12895 different datasets submitted by independent research groups as part of grant requirements, as well as the ABCD study, which includes 15312 child and adolescent fMRI datasets. These were downloaded in bulk from the NIMH servers, using a query that included every fMRI dataset available that had a corresponding structural image. The format of the label from NIMH are large CSV files that include descriptions and different covariates for large chunks of, though not all of, the data. Labeling was inconsistent across collections, but many datasets included descriptive fields specific to each study, which included key words that indicated different classifications for each study. After manually reviewing these descriptions across datasets

and identifying their use in respective studies, labeling for autism versus healthy control, resting-state, and sex was performed using a Python program that pulled key words from these descriptions. Information about age, depression, and handedness was also collected. Due to the need to conserve memory on servers, datasets that failed format conversion or preprocessing were deleted, so the actual number of datasets downloaded from the NIMH exceeds this number; additionally, data releases after 2018 were not used.

UK BioBank UK BioBank data included 27,870 task- and resting-state datasets, from healthy adults between the ages of 40 and 70. These downloaded to Cambridge servers in a separate project, though I was granted permission to use this data. As the BioBank is a centralized initiative, the data was already labeled. It consisted mostly of healthy controls, though a large number reported subclinical depression.

OpenfMRI: OpenfMRI is a repository of datasets voluntarily submitted by different research groups, often in experiments related to psychological tasks. Many individual datasets came with a corresponding CSV file that included information about age and sex, and these files sometimes included information about handedness as well. However, many of the datasets did not include any covariate information, and research into the respective publications resulting from these datasets, and even direct enquiries to the research groups that submitted these data usually revealed that this information could not be obtained. As such, much of the data in OpenfMRI was unusable. Many of the OpenfMRI datasets included multiple runs from the same individual, usually to indicate different memory tasks or timepoints; all of these were included in the classification task as separate datapoints, though, as will be specified later, measures were taken to ensure that no one subject was used in both the training and test/validation sets.

ADNI, ABIDE, and ABIDE II: The Autism Brain Imaging Data Exchange (ABIDE) I and II, the International Consortium for Brain Mapping (ICBM) and the Alzheimer's Disease Neuroimaging Initiative (ADNI) datasets constituted a relatively small portion of the data used, though ABIDE I and II were a large bulk of the non-control data used in autism classification. All of these data were downloaded from the University of Southern California Laboratory of NeuroImaging website. Covariate information was included with the data in a CSV file.

ICBM: the International Consortium for Brain Mapping (ICBM) presents a relatively large collection of control datasets, though fMRI was only collected in their UCLA site (largely

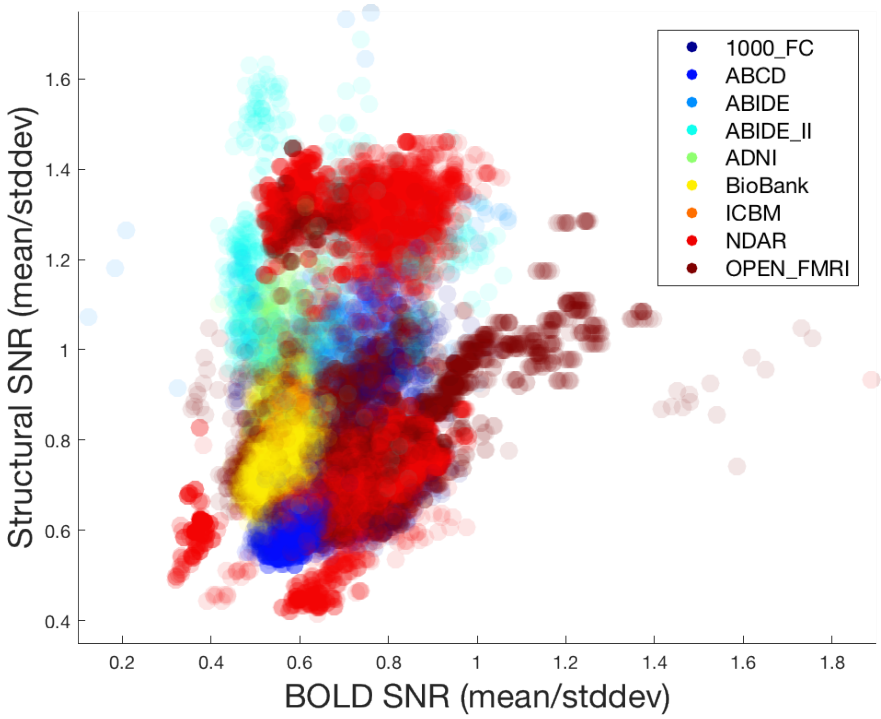


Figure 2.1: The signal-to-noise ratios (SNRs) of the raw BOLD and structural MRI files from each collection, prior to any preprocessing, taken by dividing the mean of the nonzero voxels by their standard deviation (note that this refers to SNRs from an information theory standpoint, rather than the term as it is usually used in MRI analysis). Some of the data had either extremely high BOLD or structural SNR (4 or greater), but this plot is zoomed in to display the vast majority of data.

task data) and the Montreal Neurological Institute (MNI) (largely resting-state data). From this, I used 410 task fMRI datasets from 83 unique subjects and 35 resting-state fMRI datasets from 35 unique subjects. Like ADNI and ABIDE, these were also stored on the USC database.

1000 Functional Connectomes Project: The 1000 Functional Connectomes project includes resting-state functional connectomes from various sources; much like OpenfMRI, it included covariate information contained in separate CSV files, though these were more standardized and label collection was easier.

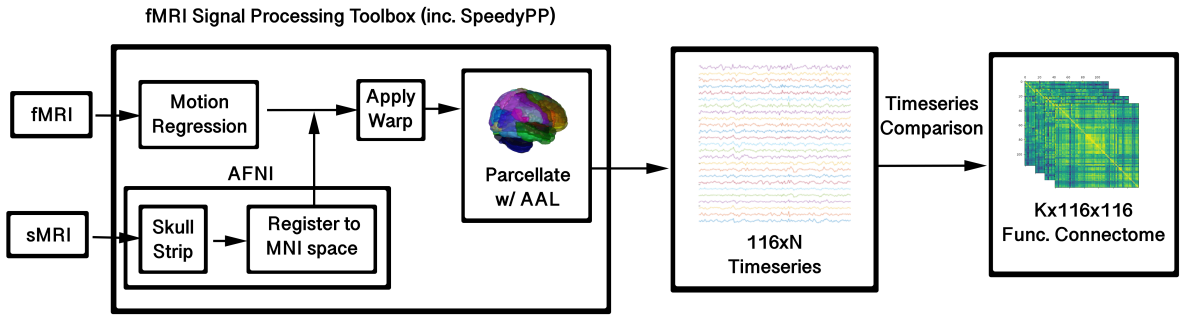


Figure 2.2: A high-level description of the functional connectivity preprocessing pipeline, from input structural and functional MRI data, to the output functional connectome. The selected timeseries comparison metric varies throughout this thesis, ranging from partial correlation, Pearson correlation, and normalized mutual information, producing a 116×116 functional connectome, to multi-band wavelet correlation, producing a $3 \times 116 \times 116$ or $4 \times 116 \times 116$ functional connectome.

2.1.2 Distribution of data quality

Figure 2.1 shows a simple calculation of the signal-to-noise ratio of the raw voxel values in both the structural and functional data (note that this does not utilize the more complex signal-to-noise calculations often applied to MRI data; this is simply a display of the mean divided by the standard deviation of nonzero voxels, as a means of showing data distribution). This displays the fundamental differences in raw data between these collections, likely due to differences in sites and preprocessing practices of different databases. Because of the differences between collections in terms of sex, rest/task, and autism, this also makes clear the necessity of balancing data by collection prior to building a training set, as Figure 2.1 shows that even a basic clustering algorithm would be capable of capturing differences between collections.

2.2 FMRI signal processing toolbox

Functional data were preprocessed using the fMRI Signal Processing Toolbox and the Brain Wavelet Toolbox (Patel et al., 2014; Patel and Bullmore, 2016) on the 116-area Automated Anatomical Labelling (AAL) atlas (Tzourio-Mazoyer et al., 2002). Following skull-stripping, motion correction was accomplished using SpeedyPP version 2.0, which utilized AFNI tools and wavelet despiking (Patel et al., 2014; Patel and Bullmore, 2016), with a low-bandpass

filter of 0.01Hz, in addition to motion and motion derivative regression. Both functional and structural datasets were non-linearly registered to Montreal Neurological Institute (MNI) space (Collins, 1998) and parcellated using the 116-area automated anatomical labeling (AAL) template (Tzourio-Mazoyer et al., 2002), which includes subcortical regions. Extracted time series were the means of each AAL region. A high-level description of this pipeline is shown in Figure 2.2. Due to the large number of data in this study, quality control was not performed during the preprocessing stage. Table 2.1 shows the percentage of datasets for each collection that passed through each stage of parcellation. Note that this table includes specialized structural measurements, which are detailed in Chapter 3. Because specific preprocessing choices differed between the presented deep learning studies, study-specific details, such as the time series comparison metrics used to create functional connectivity matrices, are given in later chapters.

2.3 Dataset counts

Exact counts of data in the context of different chapters were complicated by several factors, listed here:

- Raw datasets were downloaded from various databases, each comprising one structural and one or more functional datasets. Because one subject may have had data from multiple fMRI procedures, there were fewer structural than functional datasets. However, to simplify the storage and preprocessing of data at scale, multiple copies of structural datasets were created such that each fMRI had a corresponding structural MRI file.
- Failed data format conversions eliminated many datasets from consideration. The preprocessing of functional connectomes and structural connectomes was performed independently; thus, a dataset may have failed the preprocessing step for one but not the other.
- In order to effectively compare functional connectome classification to structural classification (as is done in Chapter 8), it was necessary to duplicate structural connectomes in order to maintain the same number of datapoints across the same unique identifiers, though this also means that the structural classification sees less data, and that may have negatively affected its accuracy. Thus, many of the structural files listed in Table 2.1 are duplicates.

- Differing uses of wavelet correlation estimations, which forced the exclusion of several datasets due to a too-low TR rate, and Pearson correlation.

- Chronology of the studies complicated the counts of data. For the most part, the studies in presented chapters were undertaken in chronological order, over a period from 2016 to 2020; the only exception to this is the structural connectivity metric presented in Chapter 3 and the visualization of vertical-filtered models presented in Chapter 7. Some data were added to their respective collection later on due to scheduled releases; in other cases, I later found errors in the original data processing scripts that led to fewer failures in format conversions (notably for NDAR and ABCD datasets).

- Different chapters had differing regional dropout criteria (i.e., in which no signal is detected when parcellating data to the AAL template). In some chapters, datasets with no more than 10% regional dropout were included, whereas in others data were excluded if any dropout was present.

As of 2020, 70,331 potential functional matrices were identified across all databases (i.e., the databases listed the functional MRI and a corresponding structural MRI as being present). Of these, 47,732 unique functional MRIs were actually downloaded, with 49,182 non-unique structural files. 25,166 total unique structural connectomes were successfully preprocessed, though accounting for duplicates across subjects, 47,359 were generated. 39,461 unique functional connectomes were successfully preprocessed. Pairing the successful connectomes with the successful structural connectomes resulted in 33,547 total; most of this admittedly-significant drop can be explained by ABCD, which dropped from 11,789 correlation matrices to 7,063 when paired with structural connectivity matrices, mainly due to regional dropout.

Exact counts of data, as well as exclusion criteria, are given in the context of future chapters as necessary.

2.4 Deep learning model

2.4.1 Implementation

The deep learning models in this thesis were all based on an implementation in Python that used the Keras deep learning library, supported by a Tensorflow backend. Keras was selected

as a machine learning library of choice due to its design, ease of use, and wide number of supporting libraries. Preliminary tests that utilized the cross-shaped convolutional neural network, BrainNetCNN (Kawahara et al., 2017), were carried out to classify functional connectivity datasets, and, indeed, the code behind this implementation was studied carefully, all presented studies used an altered architecture in Keras that used vertical filters instead. BrainNetCNN was originally designed using Caffe, an earlier deep learning library which was found to be poorly supported for Cambridge's servers without significant backend support, and the cross-shaped model, even when re-implemented in Keras (Leming and Suckling, 2019), was found to suffer from frequent training failures that disappeared when replaced with vertical filters (see Chapter 4.

All code that performed class balancing (Chapter 6) was entirely original and produced outside of the Keras framework, as well as code that scrambled input data for the stochastic models (introduced in Chapter 7). Furthermore, code to average ensemble models (introduced in Chapter 4) was implemented by using text file outputs from individual CNN models.

Code for normative pathways was implemented entirely using libraries from Matlab (see Chapter 3). Code for estimating the structural connectivity metric (also in Chapter 3) relied on a combination of FSL and original Python code.

2.4.2 Implementation of visualization methods

Activation maximization

Activation maximization (Erhan et al., 2009) is a technique to determine the maximally activated hidden units in response to the test set of the CNN layers following training. This method is presented in Chapter 5 to analyze the ways in which data bottlenecked and was organized in a convolutional layer. The code of activation maximization was entirely original, with the Keras model being edited to output maximal activations as text files; the analysis and graphing of this data was performed in MATLAB.

Occlusion

Occlusion (Zeiler and Fergus, 2013) refers to the omission of different parts of data and recording which parts of the data lead to the greatest drops in accuracy. This indirectly

estimates which parts of data are most salient for classification. Occlusion is used extensively in Chapter 7, but due to its particularly high computational requirements for the purposes of these studies, it is only seen in that context (this limitation is discussed further in Chapter 7). Occlusion was performed using entirely original Python code, with the stochastic deep learning framework also presented in Chapter 7.

Class activation maps

In Chapters 7 and 8, I deployed class activation maps (Simonyan et al., 2014; Kawahara et al., 2017; Khosla et al., 2018) using a previous Keras implementation (Kotikalapudi and contributors, 2017) to display the parts of the connectivity matrix the CNN emphasized in its classification of the test set. The original code I produced for this was to average class activation outputs from the code given in Kotikalapudi and contributors (2017).

2.5 Use of Connectivity

In this thesis, I opted to classify data based on connectivity matrices, which model covariances in data rather than data itself (Sporns, 2010). This choice was made for two primary reasons, which are explained here.

The first reason is related to formatting. Besides the practical advantages in saving memory (covered in Section 2.6), connectivity matrices maintain a consistent dimension that makes them ideal for inputting into a convolutional neural network. Raw MRIs from different sources tend to have inconsistent dimensionality; while 3-dimensional images may be resampled to a consistent dimensionality, achieving this in the fourth dimension for fMRI data or, by extension, a timeseries derived from fMRI, is nontrivial, due to wide differences in fMRI sampling rates and sequence lengths. Such data may be encoded using recurrent neural networks or LSTMs, but these are more suited to making predictions about particular timepoints rather than an entire timeseries.

The second reason is that functional brain networks, which have been the subject of intense study in recent decades, especially in relation to the phenotypic differences studied in this thesis, only emerge explicitly when covariances in data, rather than variances themselves, are analyzed. While deep learning models may infer covariances, this should not be assumed.

Additionally, modelling covariances is a form of feature extraction that maintains a form of spatial encoding in a way that variance analysis (such as voxel or ROI intensities in structural data, or t-statistics for functional data) does not.

2.6 Practical limitations

Limiting factors on research included the amount of server space, the size of working memory allowed for any one job, allotted time to run a job, and the real time it took to run multiple jobs. Certain research questions of interest were impractical to answer due to the high computational power necessary.

Across all collections, after data format conversions, a single structural MRI had an average size of 171 MB per file, while preprocessed BOLD MRI had an average size of 176 MB per file. In contrast to raw data, connectivity matrices were just under half a megabyte each ($116 \times 116 \times 4 \times 8$ bytes = 430.592 kilobytes), less than one five-hundredth the size. Preprocessing of this data using standardized pipelines, such as SpeedyPP (Patel et al., 2014), was computationally intensive, with a single dataset requiring approximately 30 minutes to run, and had to be performed across more than 60,000 datasets (even if many failed). However, these tasks could be run in parallel.

The training of deep learning models required the loading of all, or a substantial portion of, a whole dataset during training. Training on 5,000 or 50,000 raw NIFTI files would exceed the allotted memory allowed by available computational resources; even if this were not the case (and, in theory, it may be possible to read into main memory only one batch of data at a time before clearing the space), the training time required is proportional to the size of the training data. Still another prohibitive reason was that deep learning models do not give an exact accuracy, but rather an accuracy within a statistical distribution; achieving statistical power thus required the training of multiple independent deep learning models (introduced in Chapter 4), which further strained computational resources. Thus, even though raw data encoded more information of interest and may have led to higher accuracy in deep learning studies, I studied only connectivity matrices due to concerns with practicality and reproducibility.

The choice to use multiple, independent deep learning models was also influenced by the type of computing resources immediately available. The Cambridge University Department of

Psychiatry used the High Performance Hub for Clinical Informatics (HPHI), which provided CPU clusters and allowed computing times of several days. Additionally, it provided very high storage capacity. However, it did not provide graphical processing units (GPUs) for wide use, and each computing node had a set memory limit. This prohibited the use of a high-parameter machine learning model that would have required GPUs; it also prohibited a model's ability to analyze many raw datasets at once, which would have exceeded the single-job memory limit. What the resources did allow, however, was the widespread, parallel preprocessing of an extremely large dataset using pre-existing medical imaging software, as well as the utilization of many, smaller machine learning models that trained on compressed versions of the whole dataset. This allowed for was the averaging of many outputs, reducing the noise in the experiments, and the use of ensemble machine learning models, which greatly increased accuracy and allowed for the utilization of the whole dataset in a cross-validation scheme.

2.7 Hyperparameter tuning

When building a deep learning model, one has a number of hyperparameters on the model to tune. This is usually done with a grid search, in which every hyperparameter combination over a reasonable range of values is tested. Within the models considered in the present work, these may include the number of convolutional layers, the number of convolutional filters, the number of dense layers, the number of hidden units per layer, dropout percentages, initialization techniques, the slope on leaky ReLU layers, and the encoding method. Training techniques also present their own hyperparameters, including the optimization method, loss function, learning rate, weight decay, batch size, size of training/test set division, stopping criteria, and momentum. Assuming each of these variables has even three unique values that would be reasonable to test, that would imply $3^{13} > 1,500,000$ possible combinations of hyperparameters to test. Additionally, in the context of functional brain networks, there are a number of variables to consider when preprocessing the training data. Among the major considerations are the selection of parcellations, exclusion and dropout criteria, and methods of timeseries comparison. Finally, even for a particular set of hyperparameters chosen, test set accuracy naturally varies, and so 40 to 300 models ought to be trained and their accuracies averaged, before a selection can be made; given a high volume of tests, this number may have to be raised to achieve statistical significance. Another factor in this, additionally, is that certain combinations of hyperparameter values may interact to slow

down the training process and make it impractical. Neural networks with a prohibitively high number of both hidden units and layers, for instance, may take an exponentially longer time to train than networks that are only very deep or very wide.

One option may be to train each of these variables independently, making the number of tests rise linearly rather than exponentially with the number of hyperparameters. However, this would still necessitate choosing an initial set of hyperparameters, and, with N hyperparameters and an average of K choices per hyperparameter, it would still necessitate $N \times K \times 40$ tests at minimum – which, in this case, is still a very high number. Furthermore, this would fail to capture the complex interactions between variables that a grid search would be able to do.

Another option is to use commonly accepted implementation standards across deep learning. This does not necessarily promise to yield the absolute highest accuracy possible, but it is practical.

In selecting hyperparameters for this work, a combination of these approaches were conducted. Limited access to computing resources also dictated choices in these areas. While developing the code base for the models used in this thesis, grid tests on limited hyperparameters were conducted, though this was less to optimize accuracy and more to mitigate issues with vanishing and exploding gradients (the inclusion of Batch Normalization layers at a later point in development, as well as the adoption of vertical over cross-shaped filters, helped to alleviate these issues). In other cases, parameters within the same order of magnitude as those used by others were adopted, sometimes based on the scientific work and informal advice of others using similar models. Extensive tests were conducted on the original BrainNetCNN model in its Cafe implementation, prior to re-implementation in Keras; these tests revealed very few consistent improvements in accuracy could be had by varying the number of hidden units. However, while limited grid searches could reveal places to improve the stability of the model, improvements in accuracy were quite limited and spurious with a balanced dataset. As stated in Section 1.9, however, optimizing test set accuracy should not necessarily be the goal of deep learning in medical imaging.

2.8 Note about AUROC and accuracy

In this thesis, both Area Under the Receiver Operating Characteristic (AUROC, sometimes called AUC as well) and accuracy are reported. In machine learning, AUROC compares the tradeoff between the true- and false-positive classification rates, while accuracy is a direct measure of the success of a binary classification. In machine learning, AUROC is often preferred as a superior indicator of model performance because accuracy can be misleading when class proportions are unbalanced (for instance, if a training set were 99 percent class A and 1 percent class B, 99 percent accuracy could be achieved by labeling everything as class A), though accuracy is often reported because it is more intuitively understandable. Throughout this thesis, a 1:1 ratio between classes is maintained, so a gross difference between AUROC and accuracy is usually not present. However, accuracy is most often reported in machine learning literature, so a need was found to report that metric, but in practice it was also found to be more volatile than AUROC when finer comparisons were needed, so AUROC is still preferred throughout when evaluating models.

Collections	Files		GM Measurements		Func. Timeseries	Pearson Corr.	Interpolated Wavelet Correlation Bands			
	BOLD	Struct	Conn.	Vols			1	2	3	4
1000 FC	833	791	756	791	782	781	781	781	781	764
ABCD	15312	15312	7882	7903	12130	11191	11191	11191	11191	9205
ABIDE	1830	1830	1802	1829	1352	1346	895	895	895	891
ABIDE II	1174	1174	1162	1174	817	811	698	698	698	698
ADNI	263	290	284	290	262	262	262	262	262	261
BioBank	26016	27870	27498	14501	19532	19361	19361	19361	19361	16970
ICBM	387	387	347	352	383	387	381	381	381	381
NDAR	12895	12895	12070	12577	10528	10313	10313	10313	10313	8559
OPEN fMRI	9782	9782	8826	9630	7326	7189	7189	7189	7189	6655
TOTAL	68492	70331	60627	49047	53112	51641	51071	51071	51071	44384

Table 2.1: The number of files in each collection that successfully preprocessed at each stage.

Chapter 3

Brain connectivity analysis for mental conditions

In this chapter, I present two forays into connectome analysis for the characterization of mental conditions. The first, normative pathways, describes a method of pathway analysis on groups of functional connectomes. Because this particular method may only be applied to small groups of data, it was used to differentiate between groups in a small dataset ($N = 116$) of clinically depressed and non-depressed adolescents. The second, a structural similarity metric, was used to derive connectomes from T1-weighted structural MRI; this was applied to the dataset presented in Chapter 2, specifically to study brain structural differences between autistic and non-autistic controls.

3.1 Normative pathways

Functional connectivity is frequently derived from fMRI data to reduce a complex image of the brain to a graph, or "functional connectome". Often shortest-path algorithms are used to characterize and compare functional connectomes. Previous work on the identification and measurement of semimetric (shortest circuitous) pathways in the functional connectome has discovered cross-sectional differences in major depressive disorder (MDD), autism, and Alzheimer's disease. However, while measurements of shortest path length have been analyzed in functional connectomes, less work has been done to investigate the composition of the pathways themselves, or whether the edges composing pathways differ between individu-

als. Developments in this area would aid in understanding how pathways might be organized in mental disorders, and whether a consistent pattern can be found. Furthermore, studies in structural brain connectivity and other real-world graphs suggest that shortest pathways may not be as important in functional connectivity studies as previously assumed. In light of this, I present a novel measurement of the consistency of pathways across functional connectomes, and an algorithm for improvement by selecting the most frequently occurring "normative pathways" from the k shortest paths, instead of just the shortest path. I also look at this algorithm's effect on various graph measurements, using randomized matrix simulations to support the efficacy of this method and demonstrate the algorithm on the resting-state fMRI (rs-fMRI) of a group of 34 adolescent control participants. Additionally, a comparison of normative pathways is made with a group of 82 age-matched participants, diagnosed with MDD, and in doing so I find the normative pathways that are most disrupted. My results, which are carried out with estimates of connectivity derived from correlation, partial correlation, and normalized mutual information connectomes, suggest disruption to the default mode, affective, and ventral attention networks. Normative pathways, especially with partial correlation, make greater use of critical anatomical pathways through the striatum, cingulum, and the cerebellum. In summary, MDD is characterized by a disruption of normative pathways of the ventral attention network, increases in alternative pathways in the frontoparietal network in MDD, and a mixture of both in the default mode network. Additionally, within- and between-groups findings depend on the estimate of connectivity.

3.1.1 Introduction

Resting-state fMRI and connectomics

Functional Magnetic Resonance Imaging (fMRI) acquires temporal information on blood-oxygen level dependent (BOLD) signals from the human brain. Functional connectomics (Friston et al., 1993) reduces the dimensionality of these datasets to graphs (composed of nodes, representing brain areas, connected by edges) that illustrate the relationships between areas of the brain. Graph theory estimates the qualities of brain organization with measures such as centrality (or "hubness") (Sporns et al., 2007; Joyce et al., 2010; Lohmann et al., 2010; Rubinov and Sporns, 2010; Tomasi and Volkow, 2010, 2011a; Zuo et al., 2011) and community structure (or "modularity") (Traag and Bruggeman, 2009; Mucha et al., 2010; Bassett et al., 2013; Sporns and Betzel, 2016). In general, the functional connectome is characterized by high complexity (Sporns et al., 2000; Sporns, 2006), high efficiency (Buzsaki

et al., 2004), global and local synchronizability (Masuda and Aihara, 2004), and high levels of clustering with short path lengths (Hilgetag et al., 2000; Stephan et al., 2000; Bassett and Bullmore, 2006), indicating a small-world architecture (Milgram, 1967; Watts and Strogatz, 1998).

Path analysis of connectomes

Studies of average shortest path length (Gong et al., 2009; Yan et al., 2011; Lynall et al., 2010; Betzel et al., 2014) and its inverse, graph efficiency (Latora and Marchiori, 2001), have been conducted on both binarized functional (Bassett and Bullmore, 2006; Sporns et al., 2007; Wang et al., 2009; Lynall et al., 2010) and structural connectomes (Achard and Bullmore, 2007; Gong et al., 2009; Yan et al., 2011). Related to these measures are "rich clubs" (van den Heuvel and Sporns, 2011; van den Heuvel et al., 2012) that measure the tendency of nodes with high degree to be more densely connected amongst themselves than with other nodes of the connectome, which has implications for which nodes tend to be the most utilized in pathways. Like the functional connectome, an efficient, small-world structure has been shown to characterize the structural connectome (Hilgetag et al., 2000; Sporns and Zwi, 2004; Gong et al., 2009; Yan et al., 2011). Shortest-path-based node centrality measurements (such as betweenness (Freeman, 1977), regional efficiency (Latora and Marchiori, 2001; Achard and Bullmore, 2007), and closeness (Freeman, 1979)) are outlined and discussed in Sporns et al. (2007), Joyce et al. (2010), Zuo et al. (2011), and Rubinov and Sporns (2010).

The majority of connectomic analyses assume the importance of the shortest pathway, even though real-world networks often do not have knowledge of their own global structure (Boguña et al., 2009; Abdelnour et al., 2014; Goñi et al., 2013b), and so in practice, the shortest pathway is unlikely to utilized by prior planning (da Fontoura Costa and Travieso, 2007; Serrano et al., 2007; Estrada and Hatano, 2008). Studies of structural connectivity have investigated the relationships between two nodes other than the shortest pathway, such as path ensembles derived from the k shortest pathways (Avena-Koenigsberger et al., 2017), maximum flow (Yoo et al., 2015), and robustness (Kaiser et al., 2007). Furthermore, the structural connectome is both a predictor and a constraint for neural communication across the functional connectome (Passingham et al., 2002; Galán, 2008; Honey et al., 2009; Hermundstad et al., 2013; Park and Friston, 2013; Goñi et al., 2013a; Betzel et al., 2014; Mi˘ sić et al., 2015), and thus I hypothesize that alternatives to the shortest pathway provide a richer description of the topology of the functional connectome.

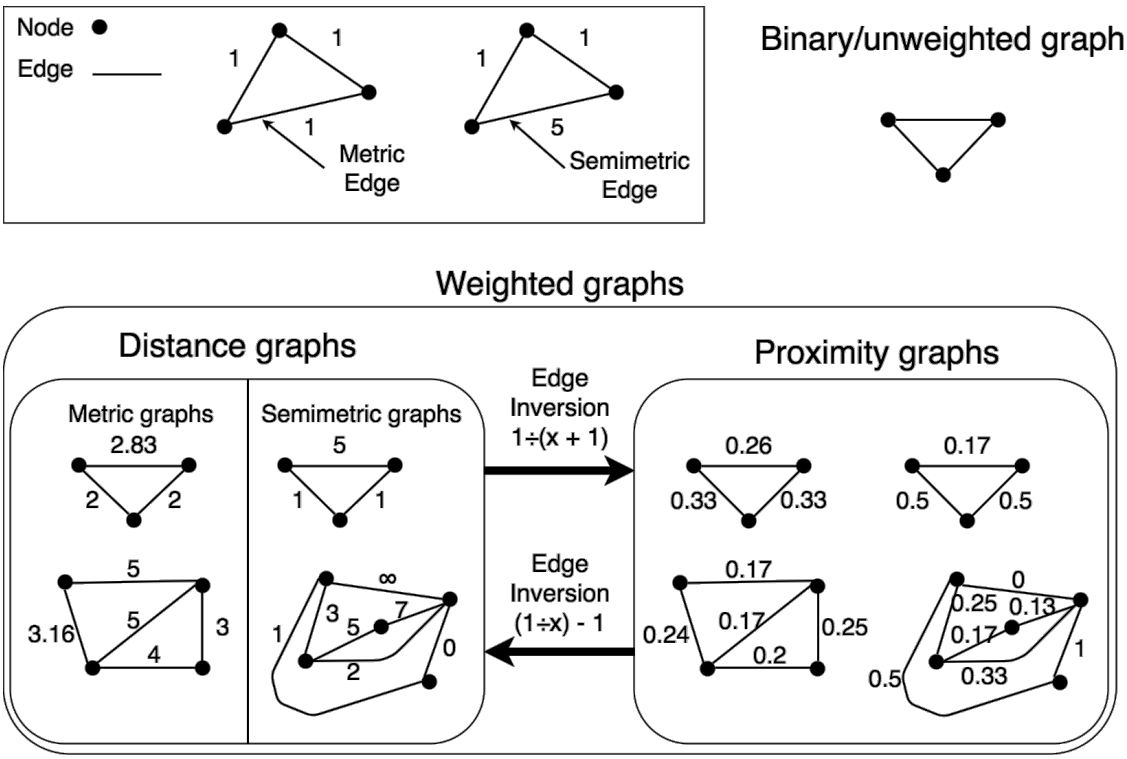


Figure 3.1: **Illustration of different graph types and the terminology used to reference them in this article.** See also Methods 3.1.2

Previous work on semimetric analysis of functional connectomes

Functional connectomes are represented, in the case of Pearson correlations, as a positive semi-definite matrix with values on [-1,1] or, in the case of alternative measurements like normalized mutual information (i.e., the shared information between two timeseries)(Kvalseth, 2017), as a matrix with values on [0,1]. It is often the case that path finding is performed after thresholding of edges to generate a binary graph with nodes defined as voxels (van den Heuvel et al., 2009) or regional parcels of the brain (Sporns et al., 2000; Stephan et al., 2000; Bassett and Bullmore, 2006). More recently, however, path finding on unthresholded functional connectomes has been undertaken (Rocha, 2002; Cao et al., 2014; Simas and Rocha, 2014; Simas et al., 2015b; Suckling et al., 2015) by inverting them from a proximity graph to a distance graph, which is embedded in a semimetric space (see Methods and Figure 3.1).

Previous studies have shown both sensitivity and specificity in differentiating control participants from individuals with autism and major depressive disorder (MDD) (Simas et al., 2015b) using the proportion of edges in semimetric distance space with a shorter indirect path: the semimetric percentage. Additionally, Suckling et al. (2015) used a similar semi-

metric analysis to classify patients with Alzheimer's Disease (AD). Although successful in distinguishing alterations in brain functional organization, these semimetric approaches rendered scalar measurements for large regions of the brain without investigating the origins of the changes, and in particular the edge composition of the constituent pathways.

Whilst there may be a difference in the proportion of shortest paths between two nodes that are indirect, there has not yet been a characterization of the routing of the shortest indirect paths, or their consistency of routing through particular areas of the brain. Furthermore, if the shortest indirect path among individuals is inconsistent, is there a second, third, or kth shortest pathway that consistently connects two areas? And do indirect paths differ in patients with mental health disorders; for example, MDD, as compared to healthy individuals?

Many psychiatric and neurological (Delbeuck et al., 2003) disorders are now being characterized from the perspective of altered or disrupted connectivity. Functional plasticity is central to the development and aging of the brain (Anderson and Thomason, 2013), its response to injury (Anderson et al., 2005), and neurodegeneration (Greenwood, 2007). Thus, an expansive analysis of the constellation of shortest paths that route information through complex brain networks is key to a deeper understanding of the information contained within the functional connectome. Below, I identify and analyze the *normative pathways*, which refer to a set of the most consistently occurring of the k shortest pathways across a group of connectomes.

3.1.2 Methods

Measurements and optimization

In this chapter, I define normative pathways and discuss a method for their detection, illustrating its performance on simulated data and *in vivo* images acquired in a case-control design. I present: (1) an index to measure the consistency of pathways between two nodes across a group of individuals—the Jaccard edge index; (2) an optimization problem that maximizes this index, thus identifying normative pathways, by analyzing the k shortest pathways between two nodes; (3) an optimization algorithm that heuristically estimates this problem, providing a practical means of finding normative pathways; (4) the behavior of the optimization algorithm and its ability to accurately identify normative pathways tested with simulated matrices where the ground truth is known; (5) a comparison of normative

pathways to shortest pathways, in terms of edge composition, centrality measures, and efficiency, in a group of control adolescents; and (6) a derivation of a statistical method for the detection of differences between normative pathways in two groups of connectomes, and apply it to a case-control comparison between adolescents with a diagnosis of MDD and control individuals.

In common with the overwhelming body of prior work in functional connectomics, the Pearson correlation of time-series extracted from two nodes is the estimate of connectivity that weights the edges between the nodes in the connectome. However, I also applied the methods of identifying and comparing normative pathways when estimating connectivity with partial correlation (which regresses out the time series of every other node in its comparison), and normalized mutual information (which quantifies the shared information between two variables). I refer to different connectivity measurements as *modalities*, and in each experiment I compare across modalities.

Terminology

A graph, G is defined as a set of nodes, or vertices V, connected by edges E ($G = V, E$), that may be directed or undirected, depending on whether edges have associated directionality. Functional connectomes are generally undirected graphs of which there are two types: weighted and unweighted (or binary) which, respectively, refer to graphs with and without numerical values associated with their edges. I use the term *proximity graph* when larger edge values represent stronger connections. Thus, in a proximity graph with edge values on $[0, 1]$, 0 represents a weak association and 1 represents a strong association. Conversely, I use the term *distance graph* when smaller edge values are associated with stronger (i.e. closer) connections and larger edge values are associated with weaker (i.e. more distant) connections. In distance graphs, edges may be *metric* or *semimetric*, depending on whether or not they satisfy the triangle inequality. Thus, an edge is semimetric if it is not the shortest path between the two nodes it directly connects. See Figure 3.1 for a visual depiction of these different terms.

Participants and MRI data

Data used in this chapter was collected as part of the MR-IMPACT study (Hagan et al., 2013, 2015). BOLD-sensitive MRI were acquired on a Siemens 3T Tim Trio scanner located

at the Wolfson Brain Imaging Centre, University of Cambridge, UK whilst participants were resting with eyes closed. Details of the MRI acquisition parameters as well as explanations for participant exclusions can be found in Chattopadhyay et al. (2017). Participants and their families gave written and informed consent, and ethical approval was provided by the Cambridgeshire Research Ethics Committee (Reference: 09-H0308-168).

The control data were taken from a sample of 34 healthy adolescents (7 males and 27 females, aged 12 to 18 years, mean age = 15.7, standard deviation = 1.45) with no family history of depression, who were recruited by advertisement from local schools. Forty (40) were initially recruited, with a total of 6 excluded. All of the participants were rescanned six months later as part of a longitudinal study, with four excluded.

Patients with MDD were recruited from East Anglia and North London, United Kingdom. 109 participants were reported in the MR-IMPACT study (Hagan et al., 2015), and of these 108 were used in Chattopadhyay et al. (2017), with exclusions for 26 participants on the basis of head motion, psychosis, withdrawals, parcellations, dropouts, and missing data, leaving 82 (18 males and 64 females, aged between 13 and 18 years, mean = 15.6 years, standard deviation = 1.12 years) for inclusion in this study.

Deriving the semimetric connectome

Due to controversies around interpretation of negative correlations between brain regions (Fox et al., 2009; Murphy et al., 2009), when constructing a graph from estimates of connectivity I first take the common step of setting the negative Pearson correlations to zero (Cao et al., 2014), and additionally set those correlations with an associated $p > 0.05$ to zero (this value is, of course, arbitrary, but it is an effort to eliminate spurious connections). To convert the edges of this weighted proximity graph to a distance graph on which path finding algorithms may be applied, I use a mathematical construct called a t-norm that converts from $[0, 1]$ to $(\infty, 0]$ using a version of the Dombi t-norm (Dombi, 1982; Simas, 2012):

$$f(x) = \frac{1}{x} - 1 \tag{3.1}$$

As these inverted weights may violate the triangle inequality, the distance graph is embedded in a semimetric space in which path finding algorithms may be applied. To convert back from semimetric distance space to a proximity space, I apply the inverse of Equation 3.1:

$$f^{-1}(x) = \frac{1}{x+1} \tag{3.2}$$

Path length measurements

Functional connectomes are most commonly represented as proximity rather than distance graphs, and thus it is convenient to also express paths in proximity space. I therefore find path lengths by first converting the graph from a proximity to distance space (Equation 3.1), summing the distances, and then converting back from distance to proximity space (Equation 3.2).

Within a distance graph, the path length is the sum of the values of edges that make up a path between two nodes. Within the Pearson correlational space that has negative correlations set to 0 and using a Dombi t-norm to sum correlations, the path, P, from node i to node j, consisting of correlations (i.e. edges) $\{P_1, P_2, ...P_n\}$ is summed to weight $W(P)$:

$$W(P) = \frac{1}{\sum_{i=1}^{n}(\frac{1}{P_i} - 1) + 1} \tag{3.3}$$

This is simply Equation 3.1 (the Dombi t-norm) embedded in a summation within Equation 3.2 (the inverse of the Dombi t-norm).

Jaccard edge index

When assessing the shortest pathways connecting nodes i and j in two functional connectomes, both may have similar lengths yet be routed through different brain regions. Thus, when comparing pathways connecting two areas across individuals, not only is the length ($W(P)$, Equation 3.3) of the paths important, but also their edgewise composition. I perform this comparison by viewing a path as a set of edges.

The Jaccard index is a value between 0 and 1 that compares the composition of two sets:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{3.4}$$

If $J(A, B) = 1$, then A and B are identical sets, and if $J(A, B) = 0$, then A and B have

no elements in common. Thus, taking the Jaccard Index of the edges of two paths gives a measure of their similarity; i.e., the number of edges the paths have in common divided by the number of unique edges that compose the two paths. Across multiple paths, the index is averaged between each pairing of pathways. For example, suppose $\eta(G, i, j)$ returns the shortest path from nodes i to j for graph G. Then, for N graphs, $[G^1, G^2, ...G^N]$, using Equation 3.3 ($W(P)$) to evaluate path lengths, this gives an array J with elements:

$$J_{ij} = \frac{2}{N(N-1)} \sum_{x=1}^{N} \sum_{y=x+1}^{N} \frac{|\eta(G^x, i, j) \cap \eta(G^y, i, j)|}{|\eta(G^x, i, j) \cup \eta(G^y, i, j)|} \tag{3.5}$$

See Figure 3.2 for an illustration of the Jaccard edge index on toy graphs.

The Jaccard edge index provides a measure of consistency of the shortest paths across a group of functional connectomes. If $J_{ij} = 1$, then the same pathway connects nodes i and j in all connectomes; if $J_{ij} = 0$, then the pathways connecting i and j do not have a common edge. To find a global measurement of shortest path consistency, I took the average of the $n \times n$ matrix J, excluding redundant paths:

$$J_{global} = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{i=j+1}^{n} J_{ij} \tag{3.6}$$

I refer to Equation 3.5 as the Jaccard edge index of the path connecting nodes i and j, and to Equation 3.6 (the average of all Jaccard Edge Indices) as the Global Jaccard edge index.

Normative pathways

For a variety of reasons, the shortest paths may not be utilized in real-world graphs. To accommodate this perspective, I found the paths across a particular group that minimized path length whilst maximizing edge sharing. I refer to these pathways as *normative*.

I define the normative pathways as the selection of pathways that maximise the Jaccard edge index over a selection of the k shortest pathways, across a group of connectomes. Informally, this means that, instead of identifying the optimally shortest pathways connecting two nodes (which may differ depending on the connectome), I identify a set of slightly suboptimal pathways that pass through similar areas. Thus, to identify the normative pathways between

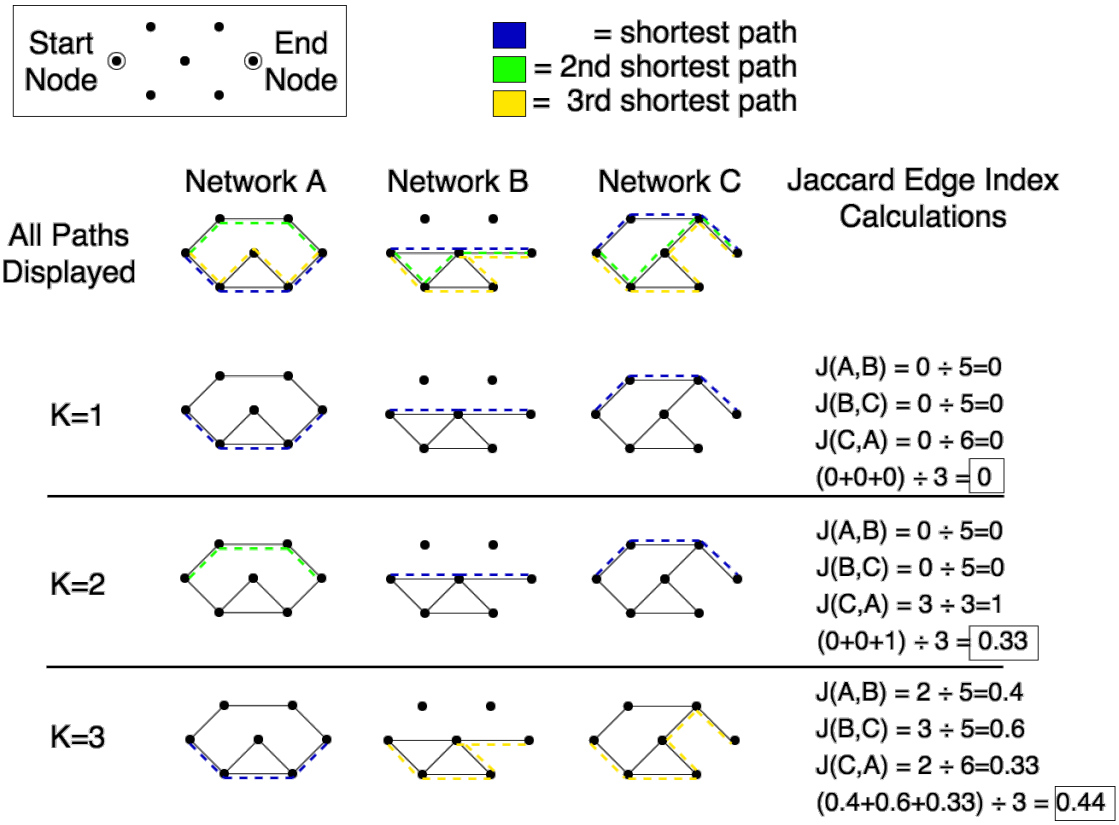


Figure 3.2: **Optimal Jaccard Edge Index that is obtained when** $K = \{1, 2, 3\}$ (i.e. when the first, second, and third shortest paths are considered) on a set of toy binary graphs. The top row displays, on the colored dotted lines, the three shortest paths of each of the three binary networks between the start and end nodes. The following three rows show which path would be selected in each of the three networks to obtain the optimal Jaccard Edge Index (if $K = 2$, the two shortest paths are considered but not the third). The function $J(x, y)$ is the Jaccard Edge index for the paths considered between two given networks. Note that the first and second shortest paths for network A are equal in length and the choice is arbitrary.

two nodes across individuals, I search for the k shortest paths that maximize J. To do this, I use Yen's K-Shortest Path algorithm (Yen, 1971), which finds the k shortest pathways by searching around each edge in the shortest path (found by Dijkstra's algorithm) and ranking the resulting paths.

Suppose that $k(G, i, j)$ returns the kth shortest path for a given connection from nodes i to j in graph G, searching across a maximum of K paths, for computational feasibility. Suppose, also, that k_l is the kth path selected for network l (so that, for instance, $[k_1, k_2, k_3...k_N] = [1, 18, 2...9]$). For a group of N graphs, $[G^1, G^2...G^N]$, the normative pathways yield the

maximum J_{ij} in the following equation:

$$J_{ij} = \max_{k_l, l \in N} \left(\frac{2}{N(N-1)} \sum_{x=1}^{N} \sum_{y=x+1}^{N} \frac{|\eta_{k_x}(G^x, i, j) \cap \eta_{k_y}(G^x, i, j)|}{|\eta_{k_x}(G^x, i, j) \cup \eta_{k_y}(G^y, i, j)|} \right) \tag{3.7}$$

An instance of the maximum Jaccard edge index found for $K = \{1, 2, 3\}$ on a set of toy graphs can be seen in Figure 3.2.

When $K = 1$ in the Jaccard edge index Maximization problem, this simply returns the *shortest paths*. As K increases, pathways become more consistent across individuals (i.e., the Jaccard edge index increases), although the path lengths become longer.

Identification of normative pathways by maximization of the Jaccard edge index

The maximisation problem expressed in Equation 3.7 is nontrivial to solve, and must be estimated via a heuristic. For each of N graphs, the k shortest paths are computed (a total of $N \times k$ paths). With each graph contributing one path, the set of N pathways is found that share the most common edges, thus maximizing the Jaccard edge index, J_{ij}.

I can maximise each J_{ij} value independently. Given N connectomes and starting from $\forall l \in N, k_l = 1$, I may iterate through $l \in N$ in random order, finding the value $k_l \in K$ that maximizes J_{ij}. I cease when no further increases in J_{ij} or can be made for $\exists l \in N : k_l < K$. I refer to this algorithm as the Jaccard edge index Maximization Algorithm.

Informally, iterating through the set of functional connectomes in random order, I test which of the k shortest paths connecting nodes i and j in a particular graph is most similar to the current set of paths from all other connectomes (Equation 3.7). I then use the path that maximizes the Jaccard edge index, stopping when no further increments can be made to the Index.

I tested the algorithm on all node pairings of a test group of 34 control participants across $K = [1...20]$, testing its ability to raise the Jaccard edge index as K increases.

Ground-truth simulation of randomized matrices

To test the efficacy of this algorithm in identifying normative pathways, I simulated randomized matrices with seeded ground-truth pathways. I randomly generated a path and seeded it into in a randomly-generated set of time series, by adding a random variable to each time series (i.e., node) that the path passed through. I varied path length (by varying the number of time series that were seeded in this way), as well as signal-to-noise ratio (by varying the weight of the random variable in relation to the time series to which it was added). I performed two classes of tests, one in which I either used a single, global random variable per simulation, and the other in which I used one per edge. To match the mean, variance, and distribution of the values in these random matrices to real data, the time series and random variables were sampled randomly from the images in the control and MDD individual's datasets. I then derived the connectomes with correlation, partial correlation, and normalized mutual information estimates and tested whether the seeded path appeared in the k shortest paths. I varied path length from 3 to 6 and signal-to-noise ratio of the simulated effects from 0 to 2 with increments of 0.025, then measured the percentage of times in 20 tests that the seeded pathway was present in the 20 shortest pathways connecting the respective nodes, for a total of 19,200 simulations. Following this, I tested whether the Jaccard edge index Maximization Algorithm converged on the simulated pathway between the two seeded nodes.

Comparing edge usage of normative pathways to that of shortest pathways

As a means of displaying which edges are more utilized between the shortest pathways and the normative pathways, I ran the Jaccard edge index Maximisation Algorithm on the test group of 34 control adolescents ($K = 20$), finding all normative pathways for each node pairing across all participants. I separately find the shortest pathways ($K = 1$). I then count, for each edge in each participant's connectome the number of times that a normative and a shortest pathway utilizes it, giving a groupwise aggregate. The counts of normative and shortest usage are each normalized to a z-score and subtracted from one another, giving each edge in the connectome a score that approximates its increased utilization by either normative or shortest pathways. Informally, this shows which areas and connections normative pathways tend to utilize more than shortest pathways.

Closeness centrality and efficiency of normative pathways

To summarise normative pathways in a connectome, I analyzed modified versions of two common graph measurements: closeness centrality (Bavelas, 1950; Freeman, 1979) and average efficiency (Latora and Marchiori, 2001). In this context, the closeness centrality of node i is the average path length ($W(P)$, Equation 3.3) of the normative paths extending from node i to all other nodes in that graph. Average efficiency is the average of all closeness centralities for a particular graph.

Both measures were modified to consider the normative ($K = [2...20]$) pathways, rather than only the shortest ($K = 1$) pathways. Given a set of paths from i to all other n nodes, $\{P_{i,1}, P_{i,2}..., P_{i,i-1}, P_{i,i+1}...P_{i,n}\}$, I define closeness centrality for node i as

$$C_i = \frac{1}{n-1} \sum_{j \in n, j \neq i} W(P_{i,j}) \tag{3.8}$$

The variance of these centralities with increasing paths ($K = 1, 2, ...20$) was also recorded, as well as the derivative of this value with respect to the number of paths used, since I am interested in the stability of these measurements as the set of shortest paths increases in number.

Additionally, the average efficiency of the graphs as K increased was calculated to observe the Jaccard edge index Maximization Algorithm's effect on global path length measurements.

$$E = \frac{1}{n} \sum_{i=1}^{n} C_i \tag{3.9}$$

I calculated these values and variances for the test and retest control groups and plotted them against K.

Cross-group normative pathway comparison

To apply these concepts to case-control studies, I look at a statistical method of comparing the normative pathways in two separate groups of connectomes, in order to detect the areas in which normative pathways converge in one group but not another and vice-versa. This method focuses on finding differences in the Jaccard Edge Indices of normative pathways

between groups that show statistically significant differences, with significance found via comparison to a null model.

For two groups, A and B, the $n \times n$ Jaccard Edge Indices, J_A and J_B respectively, are obtained for each separately:

$$J_{diff} = J_A - J_B \tag{3.10}$$

High values of elements in the resulting matrix, J_{diff}, are node pairings with normative pathways that converge to a greater in extent in Group A than Group B, while low values are node pairings with normative pathways with greater convergence in Group B, but not Group A.

To find the statistically significant values of J_{diff}, I created n null model matrices, $[J_{N_1}, J_{N_2}..., J_{N_n}]$, each found by applying the Jaccard edge index Maximization Algorithm samples of connectomes randomly assigned to each of the two groups preserving the group sizes of the observed sample. For each possible pairing, J_{diff} was calculated (Equation 3.10) to give a total of $n \times (n-1)$ different J_{diff} matrices. Subsequently, the distribution of $n \times (n-1)$ values under the null hypothesis was derived for each connection between nodes i and j. By taking the mean and standard deviation of these distributions, I converted the values of the observed J_{diff} matrix into a z-score for each node pairing:

$$J_z = \frac{J_{diff} - \langle J_{N_i} \forall i \in n \rangle}{\sqrt{\langle {J_{N_i}}^2 \forall i \in n \rangle - \langle J_{N_i} \forall i \in n \rangle^2}} \tag{3.11}$$

Z-scores, in this case, were preferred over t-scores because the population of J_{diff} matrices, growing at a polynomial rate in proporion with the sample size, is quite substantial, whereas t-scores are generally preferred for cases in which population statistics are based on smaller samples. The z-scores were converted to p-values using the Fisher Z transformation. Correction for multiple comparisons was undertaken using false discovery rate (Benjamini and Hochberg, 1995) and thresholding at $q = 0.05$, identifying the elements of J_{diff} that were statistically significant between groups.

I performed this analysis on the MDD and control groups, which gives a number of normative pathways for each group. I quantifed the number of times each edge was used across participants for each group, then determined to which regions of the brain these edges con-

nected and which they passed through most often, using the same normalization technique as above. As a means of validation, I compared these results with different studies including meta-analyses performed in MDD-control connectivity, primarily with adult participants. I counted the number of edges composing normative pathways that were significantly different that crossed through each area. The areas were then ranked and compared with those areas found to be functionally different between MDD and control adult groups in the meta-analysis of Kaiser et al. (2015). When a different parcellation, or no parcellation, was used, I manually found the closest corresponding area of the brain in the Automated Anatomical Labelling (AAL) parcellation (Tzourio-Mazoyer et al., 2002).

Code

Code for the computations was written in Matlab, using functions from the Matlab BGL toolbox (https://uk.mathworks.com/matlabcentral/fileexchange/10922-matlabbgl), the Brain Connectivity Toolbox (Rubinov and Sporns, 2010), and functions in Matlab for computing the k shortest paths (https://uk. mathworks.com/matlabcentral/fileexchange/32513-k-shortest-path-yen-s-algorithm) and the average mutual information (https://uk.mathworks.com/matlabcentral/filee average-mutual-informat). To speed up computation times, pathways were encoded as 64-bit integers, which limited the size of the pathways to $\left\lfloor \frac{log(2^{64})}{log(116)} \right\rfloor = 9$ edges for a parcellation with 116 nodes. The computations were carried out in parallel, with each graph having its 20 shortest paths for each possible connection computed independently, followed by the Jaccard edge index Optimization for $1 \leq k \leq 20$. Visualization functions were all written in Matlab, with the functions for reading in relevant NIFTI files drawing on the Vistasoft library (https://github.com/vistalab/vistasoft).

3.1.3 Results

Performance of the Jaccard edge index maximization algorithm

Figure 3.3 shows the improvements in the Jaccard edge index as $K = [1...20]$ increases, while Figure 3.4 shows the decrease in overall efficiency (indicating path length) as $K = [1...20]$ increases. The matrices for $K = [0, 10, 20]$ are shown in Figure 3.5. Each modality saw a sharp increase in internal consistency of its pathways by the application of the Jaccard edge index Maximization Algorithm, utilizing a greater distribution of edges in composition of

paths (Figure 3.6), with a small loss in overall efficiency of these paths. The most consistent pathways were seen with connectivity estimated by normalized mutual information at $K = 20$, with $J_{global} = 0.80$; in other words, the normative pathways of connectomes using the normalized mutual information modality, on the whole shared the fewest edges, but exhibited the most internal consistency.

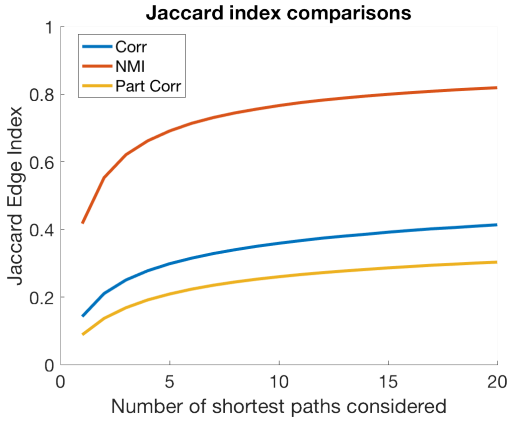


Figure 3.3: **Comparison of the Jaccard Edge Indices with normalized mutual information, partial correlation, and correlation modalities in the test group of control participants.** This displays the levels of consistency in pathways for each of the modalities. As we can see, Normalized Mutual Information offers the highest path consistency overall, being 0.80 at $K = 20$.

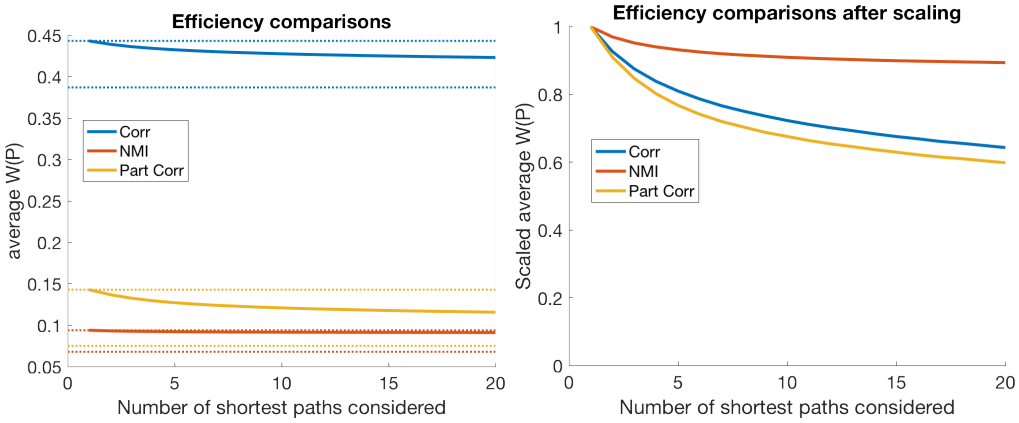


Figure 3.4: **Comparison of the average path lengths (i.e. efficiency) for different modalities over all subjects in the control group as K increases.** The left side shows the average efficiency of connectomes in the control group as more pathways are optimized for consistency; upper dotted lines represent the efficiency at $K = 1$, while the lower dotted lines represent the efficiency of the mean graph, which offers a way of scaling these lines. The right shows these three lines after scaling. See Methods 3.1.2.

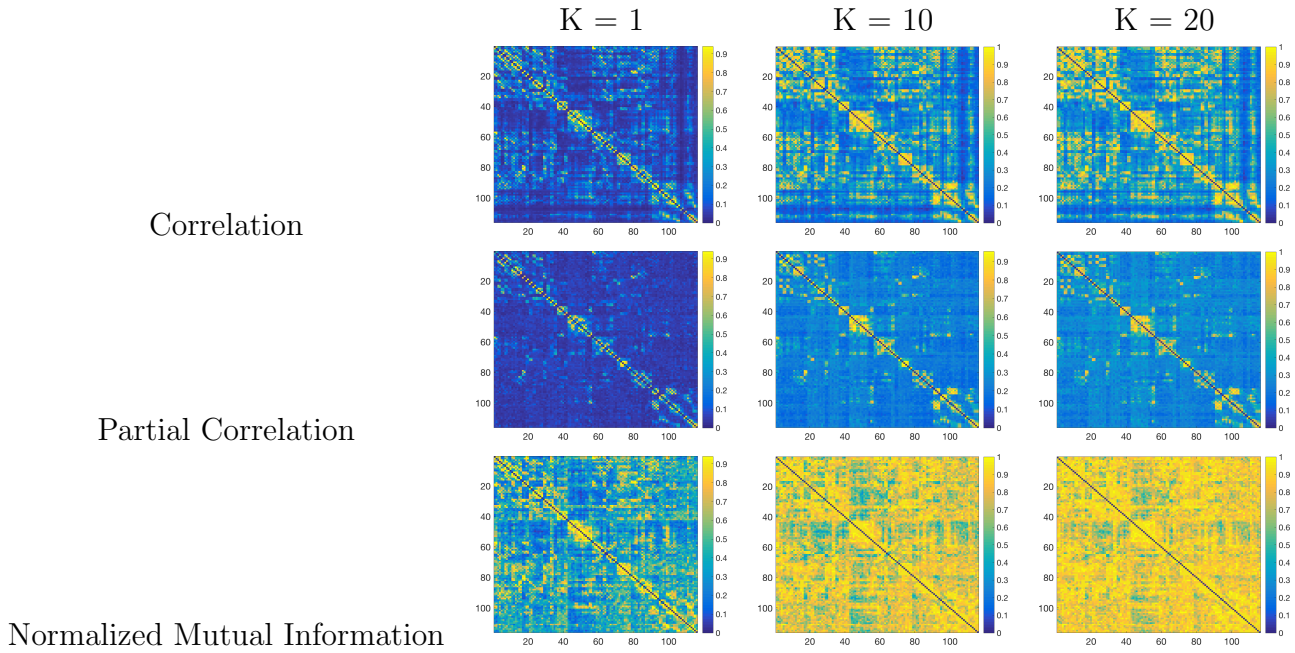


Figure 3.5: **The matrices of the Jaccard Edge Index across three modalities, only considering the shortest paths, the 10 shortest paths, and the 20 shortest paths.** These matrices compare the consistency (as measured by the Jaccard Edge Index) of the pathways selected by the Jaccard Edge Index Optimization Algorithm when considering the $K = 1$, $K = 10$, and $K = 20$ shortest pathways across a group of 30, between the 116×115 possible pathways between nodes in the AAL parcellation. The averages of these matrices over $K = [1...20]$ can be seen in Figure 3.3

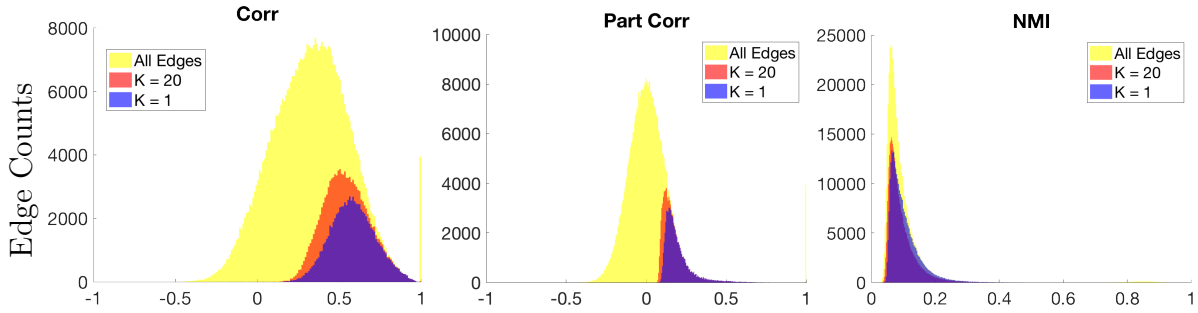


Figure 3.6: **Comparison of the overall edge utilization between the shortest pathways ($K = 1$) and the normative pathways ($K = 20$) for the 34 control subjects.**

Ground-truth simulation with randomized matrices

I simulated randomized matrices that maintained a small-world structure and degree distribution of functional connectomes by randomly sampling time series from the control and

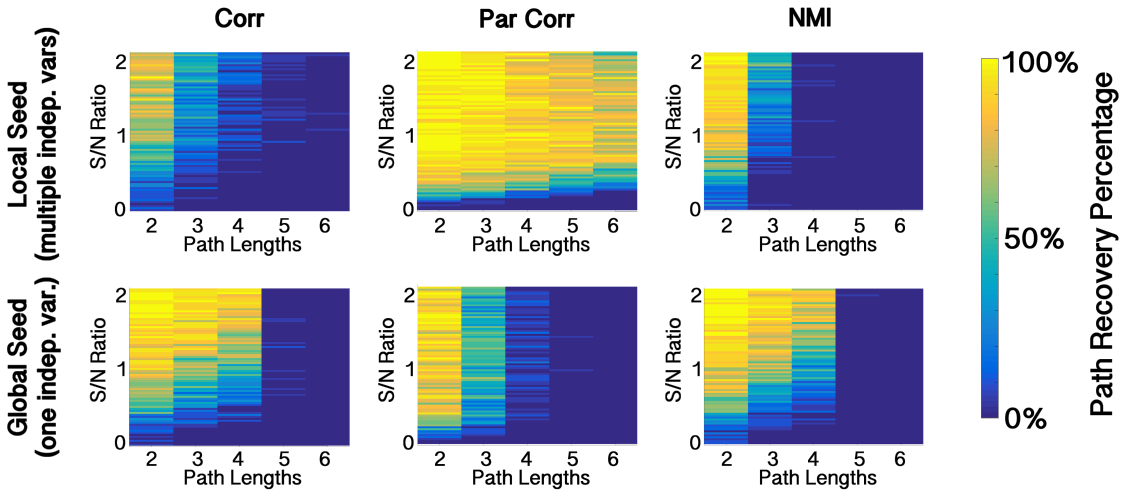


Figure 3.7: **Percentage of times that seeded pathways appeared in top 20 shortest pathways in simulated matrices.** In the top row ("Local"), one random variable was used for each edge in the path; in the bottom row ("Global"), one random variable was used for the entire path, effectively seeding a subgraph into the time series. See Methods 3.1.2.

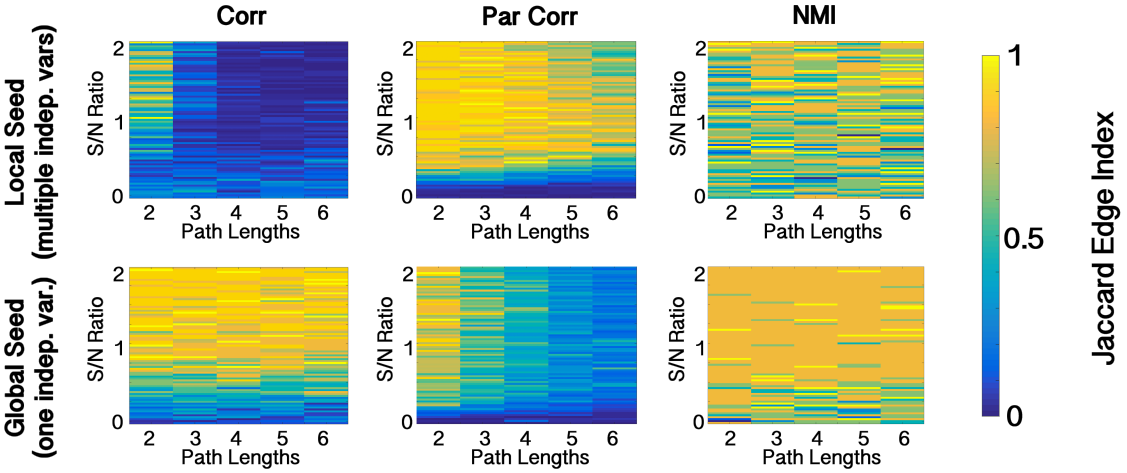


Figure 3.8: **The Jaccard Edge Indices from the simulations in Figure 3.7.** These show that, in the presence of real paths, the Jaccard Edge Index converges, even if it does not necessarily converge on the exact path that was seeded.

MDD datasets. The results are shown in Figure 3.7. I then applied the Jaccard edge index maximization algorithm to each set of 20 matrices having the same seeded path, signal-to-noise ratio, and edge independence. The means of each of the recovery percentages (the percent of tests in which the seeded path appeared in the 20 shortest paths) and the Jaccard edge index, across all tests, path lengths, and signal-to-noise ratios, are shown in Tables 3.1 and 3.2, respectively.

	Local					Global				
Path length	2	3	4	5	6	2	3	4	5	6
Corr	0.5025	0.2006	0.0549	0.0136	0.0006	0.6790	0.5469	0.4494	0.0049	0.0000
Part	0.8562	0.8210	0.7562	0.6753	0.5994	0.7216	0.3568	0.0568	0.0000	0.0000
NMI	0.6784	0.1765	0.0031	0.0000	0.0000	0.7438	0.5531	0.4130	0.0006	0.0000

Table 3.1: Mean pathway recovery percentages across all tests and signal-to-noise ratios for randomized simulations. See Figure 3.7.

When independent variables were seeded for each edge, partial correlation saw the highest success in recovering the seeded pathway, uncovering paths an average of 59.94% of the time on paths of length 6, while the highest average recovery rate for correlation and normalized mutual information for paths of length four and above was 5.49% (see Table 3.3). When a single, global variable was seeded for each path, however, partial correlation did a poorer job of recovering these pathways (as one may expect, since the single random variable, appearing in multiple time series, is regressed out), having a 5.6% average recovery rate for paths of length 4 and 0% for lengths 5 and 6. Correlation and normalized mutual information modalities had a 44.94% and 41.30% average recovery rate, respectively, for paths of length 4.

In general, the Jaccard edge index converged in the presence of a normative path with a high signal-to-noise ratio, regardless of path length. This indifference to path length is likely due to the convergence of the algorithm on another path that utilized individual edges of the seeded path. See Table 3.2 and Figure 3.8.

Due to the recovery percentages, this result indicates that the Jaccard edge index Maximization Algorithm is capable of finding seeded pathways in data, although this is dependent on both the modality and the exact method of seeding the pathways (i.e., whether I use one random variable per edge or different ones). Although this is an imperfect analogy for real-world fMRI data, it does offer an idea of the baseline efficacy of the algorithm.

Comparing edge usage of normative pathways to that of shortest pathways

Figure 3.9 shows which edges and nodes were utilized more, in aggregate, by normative pathways ($K = 20$) than shortest pathways ($K = 1$), between modalities. As expected, normative pathways utilized a wider range of edges, including weaker ones. Average path lengths of the normative pathways for Pearson's correlation, partial correlation, and normal-

	Local					Global				
No. Edges	2	3	4	5	6	2	3	4	5	6
Corr	0.3296	0.1157	0.0581	0.0596	0.0662	0.6970	0.7025	0.7048	0.6859	0.6813
Part	0.7589	0.7440	0.6999	0.6340	0.5682	0.6670	0.4925	0.3838	0.2764	0.2212
NMI	0.7978	0.8337	0.8187	0.8342	0.8300	0.8806	0.8800	0.8835	0.8761	0.8777

Table 3.2: Mean Jaccard Edge Indices across all signal-to-noise ratios for randomised simulations. See Figure 3.8.

ized mutual information modalities were 3.48, 3.93, and 2.12, respectively, compared to 3.39, 2.37, and 3.35 for the shortest pathways.

Normative pathways were more frequently routed through nodes along the upper cerebellum and the border between the brain hemispheres. With emerging evidence that white matter affects the BOLD signal in fMRI (Grajauskas et al., 2019), these are areas in which one may expect anatomical pathways to bottleneck. This is most apparent with connectomes constructed with partial correlations which showed particular increased traversing of pathways through the striatum, which receives projections from the entire cerebral cortex. Connectomes constructed with normalized mutual information showed large increases in the left and right middle cinguli; anatomically, the cingulum is a highly connected area (Hagmann et al., 2008) that acts as a global connector for other functional networks (Guimera et al., 2007; Leech and Sharp, 2014). Finally, connectomes constructed with Pearson's correlation showed large increases in parts of the upper cerebellum and vermis, and along areas directly bridging the two hemispheres; the most apparent exception, however, is between the two superior temporal lobes.

These differences suggest that normative pathways vary depending on the modality. Considering the differing values of individual edges in each connectome and the edge utilization in the $K = 1$ shortest pathways and the $K = 20$ normative pathways (Figure 3.3), this is more than likely due to inherent differences in the modalities.

Closeness centrality and efficiency of normative pathways

I measured the closeness centrality and the efficiency of normative pathways in the test group of 34 control participants for $K = [1...20]$. As noted above, the efficiency of normative pathways for all three modalities decreases as K increases, indicating that closeness centralities, on average, decrease. This is trivially true. However, I also measured the variance of the

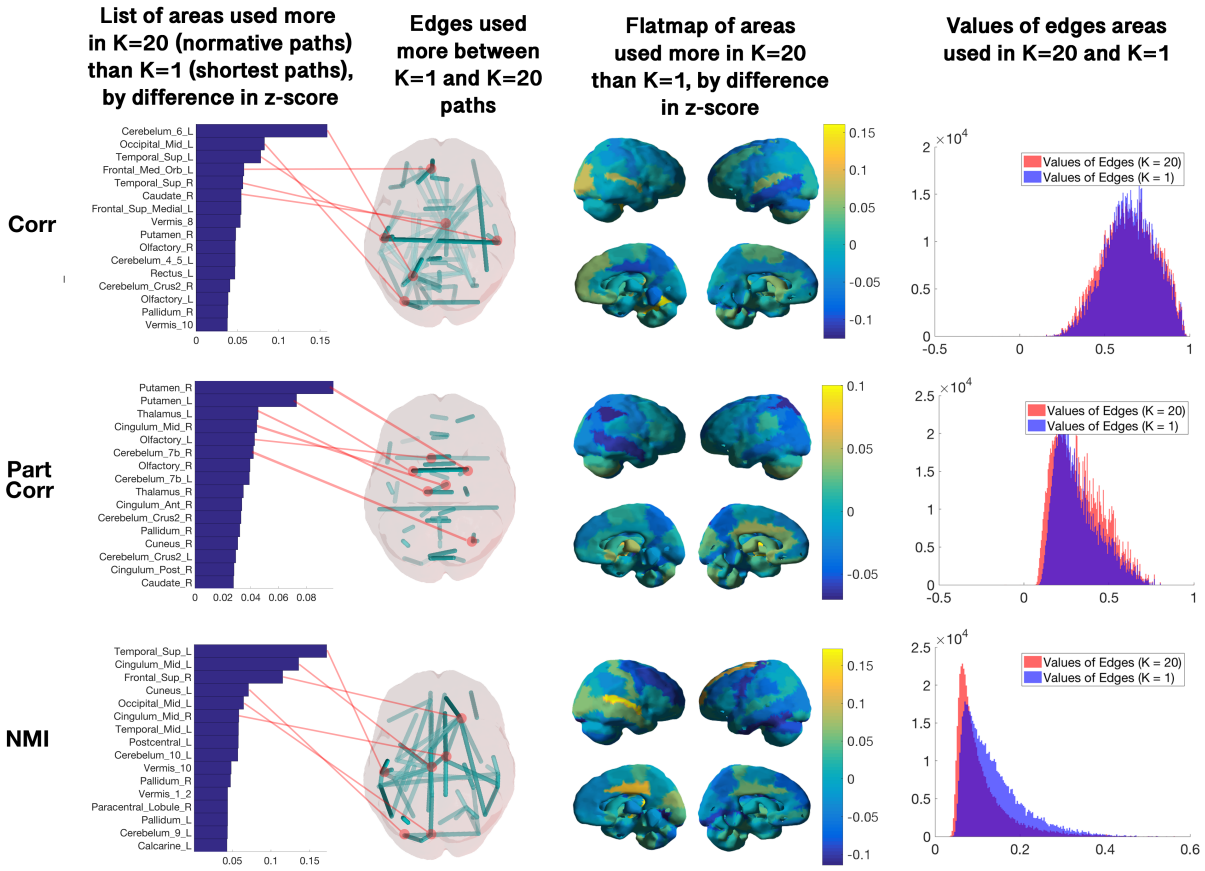


Figure 3.9: **A display of which edges are more utilized between the shortest pathways and the normative pathways in the control participants.** This visualization shows the difference in edge usage between $K = 1$ in the Jaccard Edge Index Maximization Algorithm (i.e., when only the shortest paths are considered) and when $K = 20$ (i.e., when the most consistently occurring, normative pathways are used). I counted the number of each times an edge appeared in a pathway when $K = 1$ and $K = 20$ across the 34 control participants, normalized these values to have the same mean and variance, and subtracted these normalized counts in $K = 1$ from those in $K = 20$, giving each edge a difference in z-score; the results are visualized in the second column, while the first column shows the hubs in the parcellation whose outgoing edges showed the greatest increments in utilization between $K = 1$ and $K = 20$; the flatmaps in the third column show this across the whole brain. The fourth column shows the raw values of each edge for $K = 1$ and $K = 20$.

closeness centralities across all brain areas in all participants as K increased, finding that these variances, on average, increase as K increased. The average variance of these centralities (displayed as the red lines in Figure 3.10) monotonically increased except in the case of correlation, which reached its minimum at $K = 4$ and monotonically increased thereafter.

While it may be thought that normative pathways provide more stability in their global measurements than shortest pathways, this appears to not be the case in the test group. However, in the retest group of 30 control participants (scanned six months later), the variance of all three modalities tested decreased monotonically as more paths were considered (Figure 3.11), possibly reflecting an increased stabilization effect seen on fMRI retests. Thus, correlation displays a consistent decrease in its closeness centrality variance with this algorithm for up to $K = 4$, though this is not replicated for partial correlation and normalized mutual information.

While I can conclude that the algorithm is effective in recovering normative pathways on real-world data, with substantial differences between modalities, the question of whether this algorithm has an effect on the consistency of measurements of these pathways within a group is inconclusive.

Cross-group comparisons

Using the cross-group normative pathway comparison, I performed a groupwise comparison between the control and MDD groups, finding the normative pathways that were more frequently present in one group relative to the other, across all three modalities. Figure 3.12 displays the edges most used by normative pathways that were significantly different between the test group and the MDD group. In the section below, I generalize those differences.

In all three modalities, I identified, in both groups, unique normative pathways in the frontal lobe; the MDD group had unique normative pathways in the cerebellum. Normative pathways derived from correlation and partial correlation were found more utilized in the control group in the occipital lobe. The MDD group was found to utilize a number of normative pathways more in the temporal lobe. With the exception of normalized mutual information in the control group, these normative pathways were typically local in nature, occurring within brain regions and within particular lobes.

Table 3.3 shows the brain areas connected to the most edges in each group and their associated network found to be disrupted in depression in a meta-analysis of studies with adult participants Kaiser et al. (2015). In the case of partial correlation, the clearest disrupted network intersected the occipital lobe, which has been linked to anxiety in patients with MDD (Goddard et al., 2001; Adenauer et al., 2010; Br uhl et al., 2011; Graham et al., 2013). The right cuneus and superior and mid occipital lobe were also those three areas found in

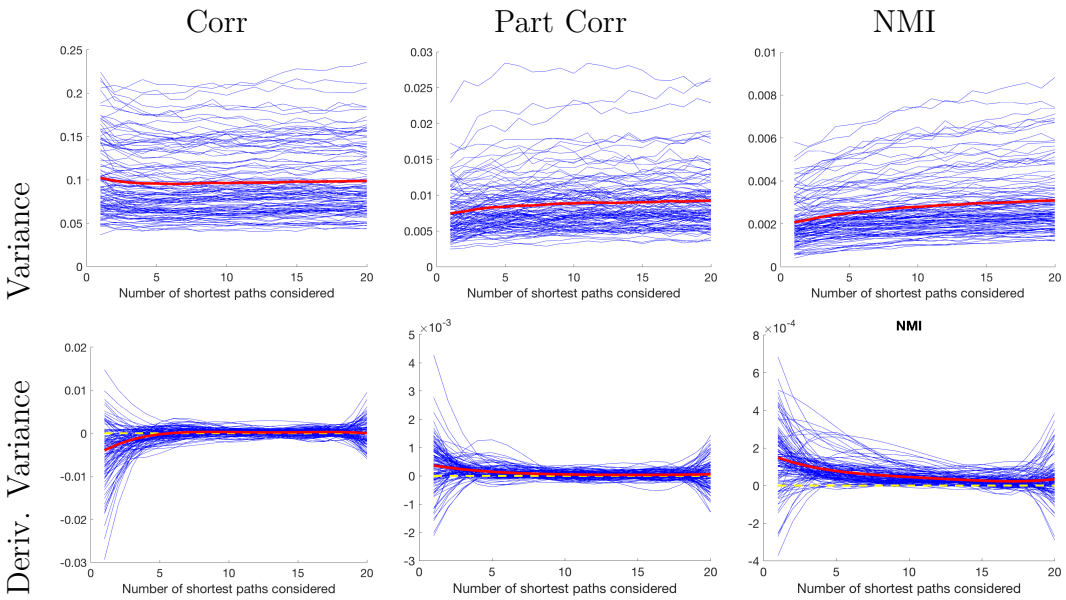


Figure 3.10: **Variance of closeness centrality for all 116 nodes in the control group.** (Above) The variance of the closeness centrality of each node, depending on the modality analyzed. Each blue line indicates the closeness centrality of one particular node in the parcellation, while the red line is the average. In general, higher variance is associated with a higher centrality value. (Below) The derivatives of these variances over path lengths considered after being fitted to a polynomial curve, which makes their fluctuations more evident. Note that, in these graphs, the order of magnitude is different, and these are meant to compare merely the fluctuations in variance per modality as the number of paths considered increases. Also note that the red lines are not true sums of variances in the true statistical sense, but are mainly used for display purposes to show general trends.

Kaiser et al. (2015) to have significant hypoconnectivity in MDD with the ventral attention network. Differential normative pathways derived from partial correlation also implicated areas previously found to be hyperconnected with the default mode network and hypoconnected to the affective network in MDD, supporting many of the findings in Kaiser et al. (2015).

Normalized mutual information and correlation detected many new and disrupted normative pathways that intersected the cerebellum. Though the cerebellum is not implicated in the meta-analysis performed by Kaiser et al. (2015), which used a multikernel density analysis, it was found to have significantly altered connectivity in rs-fcMRI in Guo et al. (2015), a study which used Pearson correlations as the estimate of connectivity, and altered negative correlations (Cao et al., 2012).

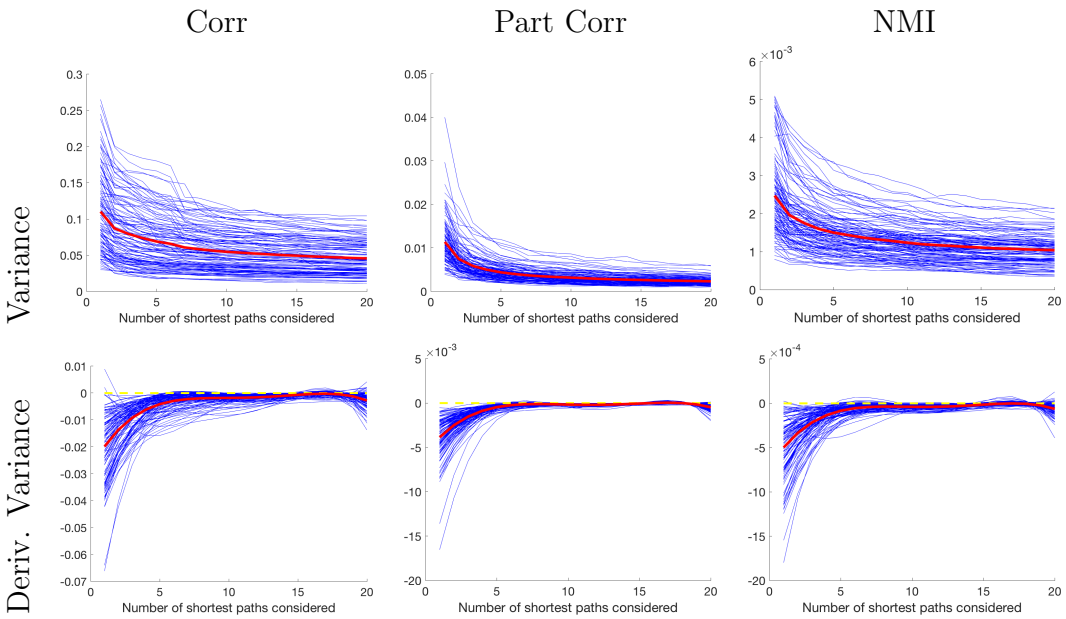


Figure 3.11: **The same closeness centrality variance results as Figure 3.10 on the retest group.** The retest group showed decreasing variances on its closeness centrality values, whereas the original group only showed this behavior in correlation.

Discussion and future work

The overarching goal of this study was to find evidence of pathways that are utilized by the functional connectome in order to discover common, potentially underlying routes of information transfer in human brains. The specific objective in this study was to find and analyze a consistent set of strong pathways in the functional connectome, to distinguish them from the shortest pathways, and to analyze the ways in which these paths differed between groups. This study provides evidence that these normative pathways are present in the functional connectome and utilized in different ways in MDD and control adolescents.

Semimetricity

The extensive development and application of graph theory in a wide range of scientific fields has encouraged its use in brain connectomics. A key concept frequently adopted is the idea of shortest pathways connecting spatially distinct regions along which information might preferentially flow. Although few studies have analysed the application of pathfinding algorithms directly in functional connectomes, such algorithms are often used indirectly; for

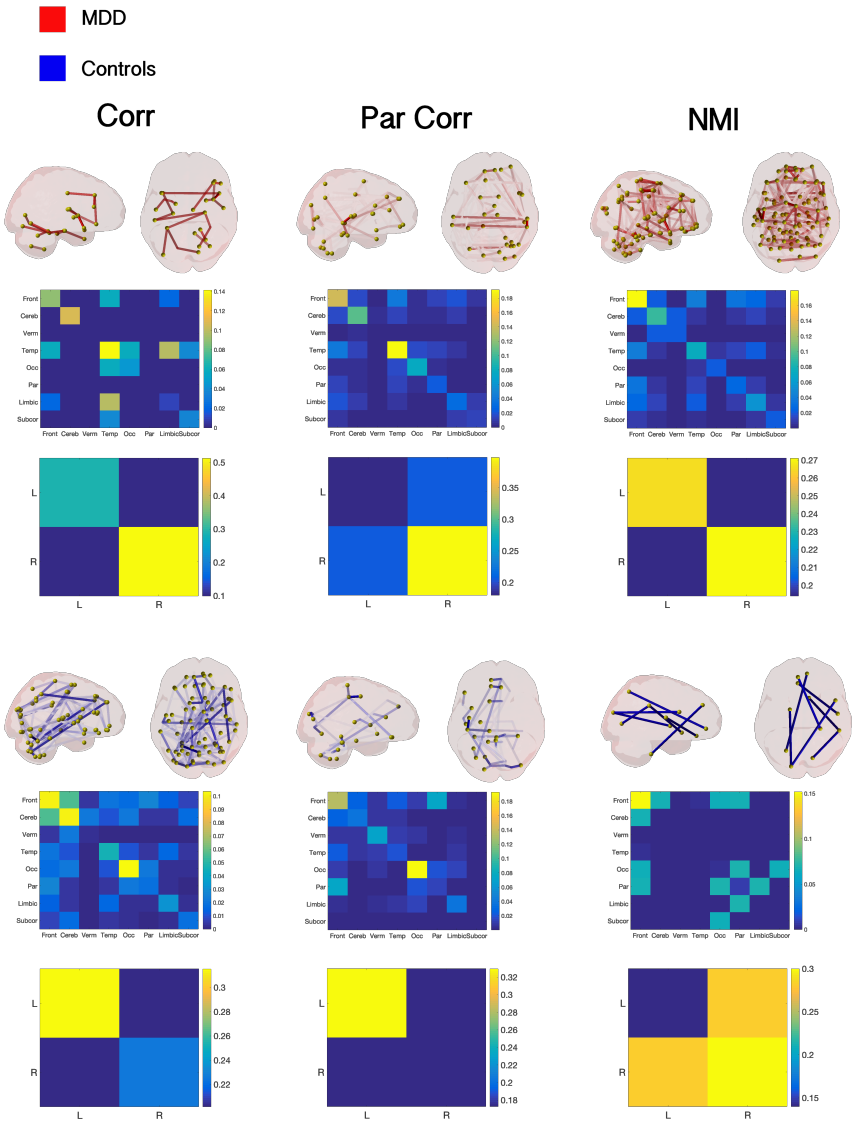


Figure 3.12: **A visualization of the normative pathways that appear differentially in each group.** These values were obtained by subtracting the Jaccard Edge Indices in each group from each other and comparing those values with differences found in a set of null models, to determine which were statistically significant. Edge intensity in the visualization is associated with that edge's use in the selected normative path in its respective group. Matrices show the fractions of edges in each extrema that connect different regions and halves of the brain.

instance, deriving system-level structures such as "rich clubs" (van den Heuvel and Sporns, 2011), and in the calculations of metrics that characterize overall network topology, such as betweenness centrality, closeness centrality (Zuo et al., 2011), and efficiency (van den Heuvel

Frontoparietal | Default Mode | Affective | Ventral Attention

Controls					
Corr		Par. Corr		NMI	
Anat. Area	# Edges	Anat. Area	# Edges	Anat. Area	# Edges
Cerebelum 6 L	1186	Postcentral L	362	SupraMarginal R	70
Occipital Sup L	626	Precentral L	208	Occipital Sup L	68
Cerebelum Crus1 L	624	Occipital Sup R	176	Cerebelum 10 R	34
Occipital Mid L	538	Cuneus R	172	Pallidum L	34
Cerebelum Crus2 R	446	Calcarine L	142	Parietal Sup L	34
Frontal Mid R	400	Vermis 1 2	132	Calcarine R	34
Temporal Sup L	348	Frontal Sup Medial R	132	Rectus L	34
Caudate R	348	Rectus R	130	Frontal Med Orb R	34
Vermis 6	312	Vermis 9	122	Frontal Sup Medial L	34
Frontal Sup Medial L	302	Rectus L	118	Olfactory R	34
Cuneus L	294	Occipital Mid R	108	Rolandic Oper R	34
Temporal Sup R	272	Cuneus L	106	Frontal Inf Oper R	34
Cingulum Mid L	264	Olfactory R	106	Frontal Sup Medial R	8
Frontal Sup R	256	Cerebelum Crus2 L	100	Cerebelum 6 R	2
Supp Motor Area R	254	Fusiform R	94	Temporal Sup R	2
Precuneus L	244	Calcarine R	86	Precuneus L	2
Frontal Inf Oper R	234	Lingual L	76	SupraMarginal L	2
Cerebelum Crus1 R	220	Paracentral Lobule L	74	Lingual L	2
Cingulum Mid R	192	Olfactory L	74	Cingulum Post R	2
Frontal Sup Medial R	192	Frontal Sup Medial L	72	Frontal Mid R	2
Cerebelum 6 R	188	Rolandic Oper L	72	–	–
MDD					
Corr		Par. Corr		NMI	
Anat. Area	# Edges	Anat. Area	# Edges	Anat. Area	# Edges
Insula R	384	Temporal Mid R	672	Supp Motor Area R	792
Temporal Pole Sup R	268	Temporal Sup R	487	Precentral L	702
Fusiform L	260	Temporal Mid L	296	Postcentral L	698
ParaHippocampal R	258	Frontal Sup Medial L	272	Cingulum Mid L	634
Frontal Mid R	240	Temporal Inf R	236	Temporal Pole Sup R	588
Fusiform R	178	Frontal Sup Medial R	228	Frontal Mid R	564
ParaHippocampal L	174	Frontal Mid R	202	Cerebelum 4 5 L	560
Lingual L	172	Cingulum Mid L	188	Rolandic Oper L	544
Temporal Inf L	164	Cerebelum Crus1 R	178	Cerebelum 6 L	524
Insula L	164	Frontal Sup R	172	Precentral R	524
Frontal Inf Orb R	164	Cerebelum Crus2 R	170	Fusiform L	514
Putamen L	150	Frontal Mid L	162	Frontal Med Orb L	494
Frontal Inf Tri R	96	Cerebelum Crus2 L	155	ParaHippocampal R	484
Temporal Pole Sup L	92	Cerebelum Crus1 L	150	Frontal Sup L	480
Cerebelum 6 R	90	Calcarine L	142	Supp Motor Area L	478
Cerebelum 4 5 R	86	Frontal Sup L	142	Insula R	470
Lingual R	86	Cingulum Mid R	140	Thalamus L	468
Cingulum Mid R	86	Caudate R	136	Cerebelum 8 L	444
Rolandic Oper R	84	Temporal Sup L	132	ParaHippocampal L	432
Cerebelum 9 R	82	Fusiform L	126	Cerebelum 6 R	426
Cerebelum Crus2 R	82	Calcarine R	122	Frontal Sup Orb R	400

Table 3.3: **Areas with the most unique normative pathways.** Shown are the 20 areas through which the most normative pathways unique to that group (see Figure 3.12) pass through. Highlighted are those anatomical areas in which differences in connectivity between adult MDD sufferers and healthy subjects were observed between the highlighted network in the meta-analysis performed by Kaiser et al. (2015).

et al., 2009). Though the idea of shortest pathways is embedded in the analysis of brain connectomes, previous studies have neither asked where these shortest pathways travel through in the brain, nor whether these pathways vary from one individual to another. Many studies have applied them to binarized functional connectomes (Bassett and Bullmore, 2006; Sporns et al., 2007; Wang et al., 2009; Lynall et al., 2010), but this approach reduces the amount of data represented by a connectome. By mapping connectomes from a proximity space to a distance space, I can apply pathfinding algorithms to weighted functional connectomes, offering a richer analysis of the data.

Normative pathways

As in other real-world networks, information does not necessarily travel along the shortest pathways (Borgatti, 2005; Hromkovic et al., 2005; da Fontoura Costa and Travieso, 2007), and indeed the quality of the information might be enhanced by additional input, I argue against the preeminence of shortest pathways in brain connectivity and suggest instead that normative pathways are a key element to the distribution of information across the connectome.

This chapter demonstrates that normative pathways are distinct from shortest pathways in the functional connectome; that inter hemispheric normative pathways closely follow direct callosal connections (Figure 3.9), which, considering previous work showing that inter-hemispheric functional connections are closely related to the integrity of the corpus callosum (Quigley et al., 2003; Johnston et al., 2008; Putnam et al., 2008; Uddin et al., 2008), suggests that they may follow the underlying biological substrate; and that analysis of their presence can yield knowledge about the differences between subnetworks in patient groups. Additionally, random matrix simulations and single-group normative pathway analysis suggest that different modalities may reveal different properties and effects in the underlying data, if they are present. This is supported by the cross-group comparison of normative pathways, which generally yielded different results depending on the modalities used, but revealed different subnetworks that were consistent with previous literature.

Studies in functional connectivity that are concerned with the analysis of the connectome itself (rather than methods of deriving the connectome from raw fMRI data, which this chapter is largely unconcerned with) are often concerned with describing the general structure of the connectome (e.g. the small-world hypothesis (Bassett and Bullmore, 2006; Sporns, 2006; Salvador et al., 2005; Achard et al., 2006)), community partitions, or finding subnetworks

such as the default mode (He et al., 2009; Smith et al., 2009; Betzel et al., 2016; Sporns and Betzel, 2016; Nicolini et al., 2017); or centrality, such as finding which parts of the brain play a central (i.e., more important) role in network dynamics (Sporns et al., 2007; Joyce et al., 2010; Zuo et al., 2011). However, initial work on functional pathways in the brain was limited due to the use of binarized networks (Bassett and Bullmore, 2006; Sporns et al., 2007; Wang et al., 2009; Lynall et al., 2010).

Avena-Koenigsberger et al. (2017) preceded this work in the use of Yen's k shortest path algorithm, arguing against the importance of shortest pathways in connectivity by analyzing path ensembles between brain regions in individual structural connectomes, relaxing the assumption that a shortest path must be taken. The presented method, by selecting one common path among a group of participants, addresses stability and reproducibility problems unique to rsfMRI (Honey et al., 2009). While this method does not exclude the hypothesis that signal communication may occur over an ensemble of pathways, it is more concerned with finding whether at least one viable pathway exists in the unstable topology of fMRI connectomes.

I found that normalized mutual information had the highest Global Jaccard edge index, correlation had the second highest, and partial correlation the lowest. This could mean that normalized mutual information naturally produces more stable pathways in its topology, or it means that other factors, such as average path length and degree distribution, trivially lower the Global Jaccard edge index. The latter reason is most likely the case. In the test group, normalized mutual information's average normative path length was smaller than that associated with other modalities (2.12 for normalized mutual information versus 3.48 and 3.93 for Pearson's correlation and partial correlation, respectively). As there are fewer possible one- or two-edged pathways that may connect two nodes, it is more likely that the Jaccard edge index Maximization Algorithm would converge on one of these as a normative pathway, thus raising the Global Jaccard edge index. Likewise, a lower average degree distribution of edge lengths (as normalized mutual information displays; see the histograms in Figure 3.9) in proximity space would inflate edge values in distance space (after application of Equation 3.1), favouring the use of fewer edges in pathways. Within modalities, the degree distributions of correlation and partial correlation were similar, with the normative pathways using a wider variety of edges than their counterpart shortest pathways (see Figures 6 and 11).

In the within-group analyses of the controls, the greatest evidence of the importance of the normative pathways is the greater use of edges along the cingulum, striatum, and the upper

cerebellum; these being central areas of the brain, one would expect them to act as bottlenecks that connect the cerebral hemispheres and the cerebellum (this is particularly true of partial correlation, which is discussed below). This supports the idea that the functional connectome is constrained by major white matter pathways, and that normative paths consisting of a larger number of edges are able to be visualized as following these constraints more closely than shortest pathways (Figure 3.9).

Modality differences

An interesting question to address is why modalities behaved in such different ways in these analyses. It is common practice in connectivity studies to select a favoured modality without considering other possibilities. This is a likely reason for some discrepancies in findings between different studies in rs-fcMRI, that can influence subsequent meta-analyses. Connectivity measures include Pearson's correlation (Eguiluz et al., 2005; Buckner et al., 2009; He et al., 2009; Wang et al., 2009), partial correlation (Liu et al., 2008; Nakamura et al., 2009; Zhang et al., 2011b), and mutual information (Salvador et al., 2008; Lynall et al., 2010; Eqlimi et al., 2013), as well as coherence (Bassett and Bullmore, 2006; Bassett et al., 2013), wavelet-based methods (Lynall et al., 2010), and other original methods that explore relationships in the frequency domain (Salvador et al., 2008; Goelman et al., 2017). Different types of analyses may also produce different results; while Kaiser et al. (2015) used a multilevel kernel density analysis, for instance, Mulders et al. (2015) looked at studies that used both a seed-based correlation analysis and independent component analysis.

I offer an explanation that partial correlation, by regressing out the global signal, is more focal in nature, and that correlation and normalized mutual information are more suited to detect global normative pathways and disruptions. Through fMRI simulations, Smith et al. (2011) found partial correlation to be among the most effective measures of connectivity, as it correctly detected connections in simulated data at a higher rate than most other modalities tested. My findings, likewise, support the efficacy of partial correlation in three contexts.

First, in the ground-truth simulations, partial correlation for locally-seeded edges detected seeded pathways more effectively than any other measurement (Figure 3.11).

Second, the within-group analysis revealed a higher path usage that included the cingulum and other callosal areas bridging the cerebral hemispheres (Figure 3.9), while the cross-group analysis implicated many areas in the middle of the cerebellum and the corpus callosum.

Previous studies have shown that the integrity of the corpus callosum is related to inter hemispheric resting-state functional connectivity (Quigley et al., 2003; Putnam et al., 2008; Johnston et al., 2008; Uddin et al., 2008). If white matter pathways bottleneck between the left and right brain in the corpus callosum, then the increased usage and emphasis of those areas is evidence that the normative pathways follow an underlying anatomical substrate that intersects these areas more frequently than the shortest pathways.

The last place that supports the strength of partial correlation is in the cross-group analysis, in which more areas previously implicated in MDD-control group differences in Kaiser et al. (2015) were found by partial correlation than the other two other modalities (Table 3.3). The Kaiser et al. (2015) analysis, however, considered older age groups, so my analysis may only have found the areas implicated in the early stages of depression.

While this indicates that partial correlation is a particularly effective means of modelling the data, I view correlation and normalized mutual information as simply alternate means of modelling the data. While partial correlation showed a clear dominance of areas adjacent to the corpus callosum, normalized mutual information also had a strong increase in usage of the left and right middle cinguli, while correlation showed increased use of edges in areas that connected either halves of the cerebellum to the rest of the brain (Figure 3.9).

When making a practical choice of which modality to use, I would generally recommend the use of partial correlation for the above reasons. Nonetheless, Pearson's correlation remains the more prevalent metric of connectivity, and its use allows easier comparison a wider variety of other studiest. Furthermore, partial correlation may be impractical on finer parcellations, or on datasets with fewer timepoints, since the number of time points cannot exceed the number of nodes in the parcellation. Normalized mutual information is advantageous in avoiding the negative edge problem.

Case-control differences in depression

Many different methods have been developed to analyze functional connectivity (Li et al., 2009). This and other studies have found many different approaches of finding groupwise differences in brain images, and these different methods often offer different results. MDD has been studied extensively (Zhang et al., 2011a; Bora et al., 2013; Graham et al., 2013; Li et al., 2013; Roiser and Sahakian, 2013; Singh and Gotlib, 2014; Qiu et al., 2015). Using different methodologies, different studies and meta-analyses have implicated case-control differences

(both in terms of structure and function) in many different parts of the brain (Kaiser et al., 2015; Mulders et al., 2015), and others have shown only limited areas of difference (Bora et al., 2013). There are several possible explanations for this. The first is that MDD is a complex disorder and each methodology uniquely captures a different aspect of the disorder. The second is that many methods used potentially capture spurious differences in the data. The third is that MDD is a system-wide disorder and different methods implicate specific parts of the brain, each partially illuminating a deeper, more widespread effect. Another explanation for the dissimilarities is the slight differences in the datasets studied; for instance, here I studied adolescents, and so a comparison to studies on MDD in adults is not one-on-one; or, individual datasets may simply be too small to give statistically reliable results. This begs the question of whether normative pathway analysis is a comprehensive means of describing a system-wide disorder, or just another analysis method that offers its view of depression.

Considerations in the interpretation of normative pathways

There are several controversies surrounding the interpretation of pathways in functional connectomes, which partially stems from controversies with functional connectivity itself. First, there are functional connections that are not fully accounted for by the underlying structural connectivity (Honey et al., 2009; Meier et al., 2016), and which may not be explained by two-edged indirect pathways. Although there is evidence in time-lag-based analyses that information propogates, either directly or indirectly, across functional connections (Cole et al., 2016; Mitra and Raichle, 2016; Ito et al., 2017), there remains concern that observed causality in the BOLD signal is due to the kinetics of neurovascular coupling (Handwerker et al., 2004; Friston, 2009).

More fundamentally, the analysis of pathways in functional connectomes is complicated by the presence of relatively strong edges that may be the byproduct of the shared variance of an indirect pathway, rather than a true instance of information transfer. For instance, an indirect pathway $B \rightarrow A \rightarrow C$ may introduce shared variance between B and C by their relationships with A that expresses as a strong edge $B \rightarrow C$, despite the lack of any direct information transfer between regions B and C.

There are means of calculating whether, for an indirect path, the shared variance between two areas is stronger than the calculated indirect pathway. Consider the three-node case, $B \rightarrow A \rightarrow C$. For Pearson's correlation, the following inequality holds:

If $corr(A, B) = c$, $corr(B, C) = a$, and $corr(C, A) = b$, then $a >= b \times c - \sqrt{1 - c^2}\sqrt{1 - b^2}$

And, if the following is true: $\frac{1}{(1+((1-\frac{1}{b})+(1-\frac{1}{c})))} > a$,

then the indirect pathway is stronger than its shared variance, and, when calculating normative pathways, the indirect pathway would rank higher in the list of the k shortest pathways than the direct edge, making it more likely that the Jaccard edge index Maximization Algorithm would converge on the indirect pathway.

To generalize this, of course, one must consider degree distributions (which, as I have shown, vary substantially between Pearson's correlation, partial correlation, and normalized mutual information), the transitivity qualities of the considered modality (i.e., the above equation for Pearson's correlation), the selected t-norm used to invert and sum edges, and the number of edges in a given path.

In general, if the indirect pathway, calculated by Equation 3.3, is consistently stronger than the direct edge connecting two areas, it is more likely to be converged upon by the Jaccard edge index Maximization Algorithm and identified as a normative pathway.

3.1.4 Conclusion

In this study, I proposed an alternative measurement to shortest pathways in weighted functional connectomes. I demonstrated that the composition of shortest pathways in functional connectomes is inconsistent and I propose a means of improving this by discovering the normative pathways. I showed that the resulting pathways from this algorithm closely utilise key anatomical areas close to the corpus callosum, which have been shown to be key to inter hemispheric functional connectivity, especially when the connectome is modelled using partial correlation. I demonstrated, as well, that the areas in the functional connectome where these normative pathways converge differently in participants with MDD and controls correspond to findings in other studies of connectivity. This demonstrates the usefulness of the method. Future studies assessing the relationship between normative pathways and underlying white matter connectivity are of importance and may improve understanding of the relationship between functional and structural connectomes.

3.2 A novel structural connectivity metric

3.2.1 Introduction

Machine learning has found multiple applications to the analysis of brain images in recent years, including pre-processing, segmentation, and diagnostics. Of great interest has been whole-brain phenotypic classification, in which MRI data of two or more phenotypes (such as sexes, or a diseased group and healthy controls) are trained and classified with a machine learning algorithm. Such studies most often include four steps: (1) selection of MRI modality and derived features that are sensitive to the problem at hand ; (2) feature extraction, to reduce data dimensionality; (3) inputting features to train a machine learning model with the selected architecture; and (4) classification and interpretation.

MRI feature extraction is most often performed using techniques previously developed in image analysis, and the specific method is dependent on the selected modality and features. For instance, based on a large body of research and predictable dimensionality reduction (Behrens et al., 2007; Kriston, 2011), it is common to use for classification functional connectivity matrices (Meszlényi et al., 2017; Kazeminejad and Sotero, 2019; Al-Zubaidi et al., 2019) representing correlations in time-series between pre-defined regions derived from blood oxygenation level-dependent (BOLD) sensitive fMRI. Likewise, to classify diffusion weighted images (DWI) it is common to use structural connectivity matrices representing the number of white matter tracts traversing the brain between specific regions (Dodonova et al., 2016; Kawahara et al., 2017; Frau-Pascual et al., 2019).

However, while there exists several consensus methods for deriving connectivities from fMRI (Kriston, 2011; Patel and Bullmore, 2016) and DWI (Behrens et al., 2007) (though this is still an active area of research (Seidlitz et al., 2018; Paquola et al., 2019)), analogous means of connectivity-based dimensionality reduction for T1-weighted structural MRI (Kong et al., 2014, 2015) are less widely used, even though this is the most common (Preston, 2006) modality available to study. One reason for the lack of common methodology is that reductions from three-dimensional data to network representations with meaningful physiological interpretation are more difficult to produce than reductions of four-dimensional data. In most existing feature extraction methods for T1-weighted MRI, extracted features are typically independent, univariate measures from regions of interest, such as cortical thickness and surface curvature. However, the lack of a connectivity metric leads not only to the loss of spatial encoding seen in network representations, but fewer features overall (i.e.,

					Age				Sex		
Collection	Subjs	FCs	Rest	Task	Min	Max	Mean	Stdev	Female	Male	Autism
ABCD	1049	5142	2296	2846	0.42	11.08	10.12	0.69	2474	2668	61
ABIDE	412	412	412	0	6.00	45.00	17.00	7.16	45	367	181
ABIDE II	682	717	717	0	5.22	55.00	14.39	7.39	169	548	350
BioBank	9791	9791	9791	0	40.00	70.00	55.00	7.51	5178	4613	4
NDAR	1050	7958	5531	2427	0.58	55.83	18.71	7.80	3816	4142	930
Open fMRI	1194	5268	820	4448	5.89	78.00	27.12	10.24	2346	2479	29
Total	14178	29288	19567	9721	0.42	78.00	30.72	–	14028	14817	1555

Table 3.4: Statistics for each dataset used.

for N ROIs, connectivities output $O(N^2)$ features while univariate measurements output $O(N)$), reducing effectiveness for machine learning.

For this chapter, I designed a similarity metric that reduced T1-weighted MRIs to a network representation without an a priori physiological interpretation, then applied it a dataset of autistic individuals and neurotypical controls. I applied this method to an extremely large dataset of participants with autism, representing a disorder for which structural characterization had proven difficult (Plitt et al., 2015; Katuwal et al., 2015; Heinsfeld et al., 2018; Khosla et al., 2018).

3.2.2 Methods

In the present work, I present a simple method of deriving structural connectivity matrices from T1-weighted MRI. My method compared the distributions of grey matter in pairs of parcellated areas of T1-weighted MRI. While this method has no specific physiological interpretation, it acted as an effective means of dimensionality reduction that allowed for T1-weighted MRIs to be encoded into a machine learning model.

Dataset

I used a dataset containing 29,288 total instances each with a structural MRI and a functional MRI in both task-activated and task-absent (rest) conditions. (Note that in many instances, data were acquired from the same participant.) In total, 1555 data points were from participants with autism. These data were drawn from six different databases: OPEN

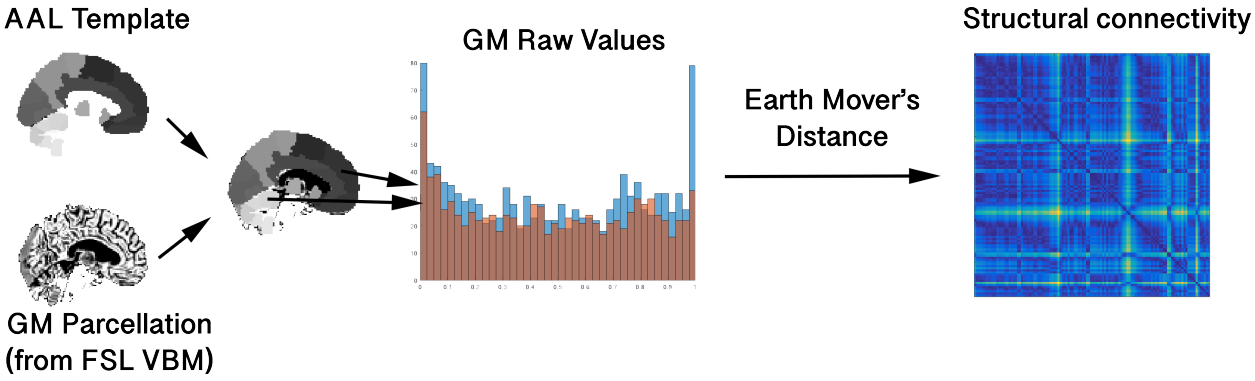


Figure 3.13: Illustration of the procedure used to estimate the structural connectivity matrices used in the present study.

fMRI, the UK BioBank, ABIDE I, ABIDE II, NDAR (minus ABCD), and ABCD (Table 3.4). Covariates of age, sex, task were also compiled.

Single-participant structural connectivity matrices

To estimate gray matter distributions in each area in the AAL parcellation, structural MRI were first skull stripped using tools from the Analysis of Functional Neuroimages (AFNI) toolbox, then registered to MNI space and grey matter values estimated using FSL VBM. I measured the similarity, s between two regions by comparing the distributions of nonzero voxel values within the distributions of each region (u and v), using the following equation:

$$s = \inf_{\pi \in \Gamma(u,v)} \int_{x,y \in \mathbb{R} \times \mathbb{R}} |x - y| d\pi(x, y) \tag{3.12}$$

in which $\Gamma(u, v)$ is the set of distributions on $\mathbb{R} \times \mathbb{R}$ whose marginals are u and v (Ramdas et al., 2017). This is simply the Wasserstein metric; intuitively, this indicates the minimal amount of work necessary to transport one distribution to another (in describing this metric, the two different distributions are often described as piles of dirt – hence its alternative name, "Earth-Mover's distance"). This is an ideal metric as it non-parametically compares two statistical distributions, regardless of relative region sizes. A similar metric, the Kullback-Leibler divergence, has previously been used in brain morphology comparisons (Kong et al., 2014, 2015), but this metric required the estimation of a probability density function rather than operating on discrete data directly, because it is sensitive to histogram binning, whereas the Wasserstein distance is less so (Rubner et al., 2000). While this similarity metric does do

away with spatial encoding and thus eliminates crucial information such as curvature, it acts as a comparison of the distributions of grey matter volumes between two areas in an easily understood way, and at a low computational cost. An illustration of this is shown in Figure 3.13. While this is a similarity metric that implies no unique physiological relationship between areas, I refer to it as a form of "connectivity" in line with the commonly used vocabulary in connectomics.

Comparison of functional and structural connectivities

To determine whether functional connectivity and the novel structural connectivity metric shared information, I correlated Pearson functional connectivity matrices from each instance with their corresponding structural matrices, in 10,000 random samples. I then compared these correlations with a null model estimated by correlating random pairings of functional and structural matrices across the collection. This comparison was done by comparing the two sets of 10,000 R values with a t-test, and indicates the amount of common information encoded by both functional and structural connectivities.

3.2.3 Results

Comparison of functional and structural connectivities

Figure 3.14 shows the average functional and structural connectivity matrices for a balanced group of autism and neurotypical controls. Correlations of functional and structural connectivity matrices from the same participants suggest a modest negative correlation. Across 10,000 random comparisons, the average R value of correlated raw edge values was -0.118 against a null model of -0.108. Subsequent t-tests showed that the R values of the direct comparisons and null model test fell under different distributions ($p=2.216 \times 10^{-13}$). This indicates that structural and functional connectivities share only a modest amount of similar information for the same participant.

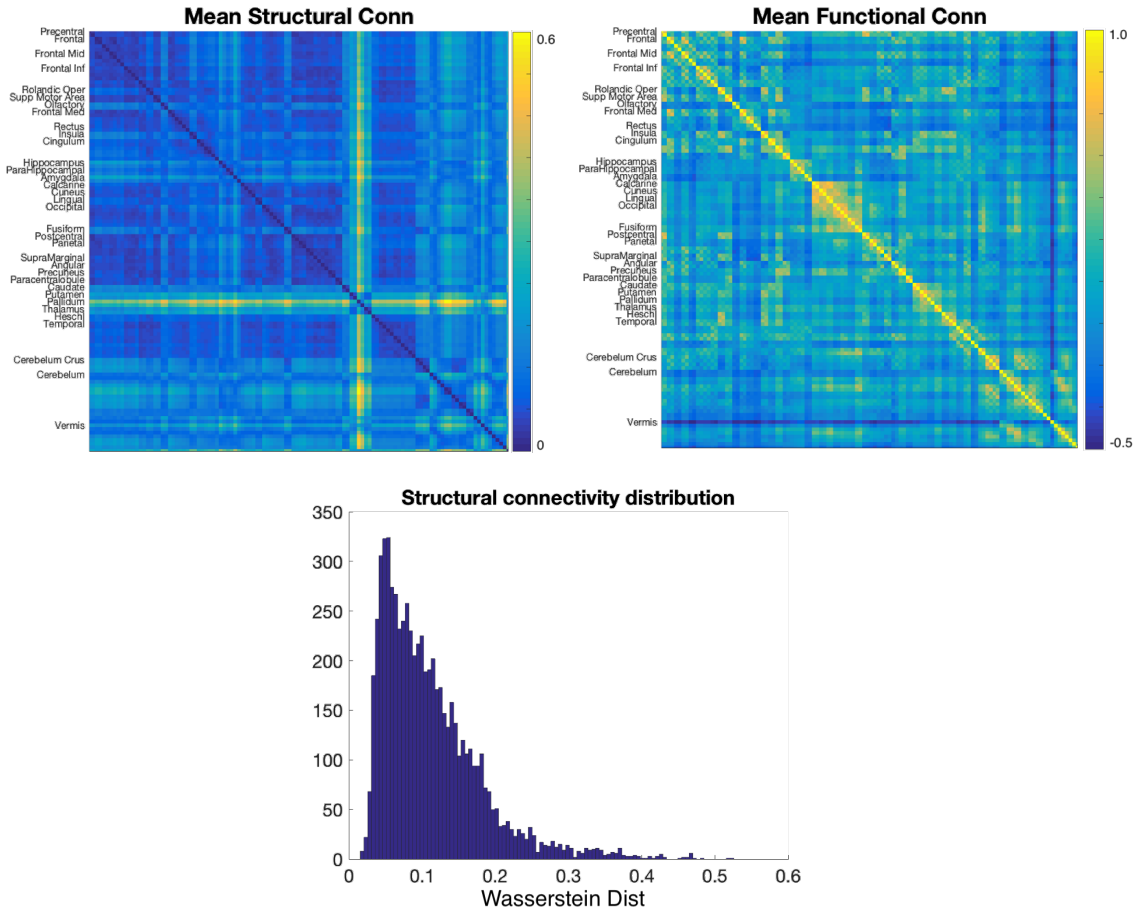


Figure 3.14: The average structural (left) and functional (right) connectivity matrices. The distribution of values of the structural connectivity metric is also shown.

3.2.4 Discussion

In this chapter, I proposed a new feature extraction method for inputting structural MRIs into a network-based machine learning model, as well as applicable analysis methods to detect areas of that were particularly involved in determining the classification. Estimating single-participant structural connectivity matrices from T1-weighted images without supplementary modalities such as DWI or fMRI is uncommon, and research in this area is ongoing (Tijms et al., 2012; Kong et al., 2014, 2015). In structural covariance, VBM data is used to produce inter-regional relationships at a group level, but this is inapplicable at a single-participant level, which is necessary to make structural MRIs applicable to machine learning models. The proposed method provides a means of doing so.

In developing this method, other means of estimating single-participant connectivity ma-

trices from T1-weighted MRI were considered, such as estimating the correlation between different univariate measurements (cortical thickness, curvatures, and so on) of the structural image (Seidlitz et al., 2018; Paquola et al., 2019), but this was too computationally intensive for a large dataset. Another method was investigated that involved finding the difference between group structural covariance matrices with and without a certain participant. While classifications on these matrices were successful, the matrices themselves varied to such an extent that the output CAMs were inconsistent. In the end, the proposed method was used because of its simplicity and effectiveness in classification.

Chapter 4

Ensemble CNNs for connectivity classification

In the this chapter, I use labels from the large functional connectivity dataset described in Chapter 2 in a deep learning task, using a model inspired by a previously-designed deep learning framework, BrainNetCNN. This study has three important innovations: (1) the encoding technique of BrainNetCNN is changed from cross-shaped to vertical convolutional filters; (2) an ensemble of models, rather than a single one, is used, to ensure statistical robustness; and (3) multi-band wavelet correlation is used, allowing me to take advantage of depth encoding built in to neural network libraries. Results from these studies, and further analysis of the ensemble, are presented. Three labels are classified: autism, sex, and resting state/task fMRI. Employing class-balancing to build a training set, I trained 3×300 modified CNNs in an ensemble model to classify fMRI connectivity matrices with overall AUROCs of 0.6774, 0.7680, and 0.9222 for autism vs typically-developing (TD) controls, sex, and task vs rest, respectively. Projections of AUROCs if models were added to the ensemble *ad infinitum* are also provided. This study is presented primarily as an analysis of autism.

4.1 Introduction

The characterization of brain differences in autism is an ongoing challenge. Although the consensus is that there are widespread structural and functional differences, the direction and spatial patterns of differences are not reliably observed and overlap with inter-individual

variability in the neurotypical population.

Estimates of grey matter volume with voxel-based morphometry (VBM) have been the most commonly used methodology to assess brain structure, but have resulted in discrepancies amongst meta-analytic findings (Cauda et al., 2011; DeRamus and Kana, 2015; Yang and Hofmann, 2015), at least a partial explanation for which are the small sample sizes that are a prevalent feature of the primary literature (Button et al., 2013; Nord et al., 2017).

To address variations in data acquisition and processing that make between-study comparisons less powerful, publicly available large-sample datasets are now pivotal to imaging research. ABIDE has made available over 2000 images in two releases, but cross-sectional VBM analyses have failed to observe significant differences (Haar et al., 2016; Zhang et al., 2018). Other morphological properties of the cortex may yield greater sensitivity (Khundrakpam et al., 2017), and recent findings using estimates of cortical thickness from the ENIGMA working group suggest a complex pattern of differences relative to neurotypical controls that varies across the lifespan (van Rooij et al., 2017). Other databases, such as the National Database for Autism Research (NDAR) act as aggregates of MRI data for different smaller-scale studies, though centre differences complicate conventional analyses on these data as a whole.

Autism has been consistently associated with differences in brain function (Müller et al., 2008; Simas et al., 2015a). This is often studied in the context of EEG (Ahmadlou et al., 2010, 2012; Bhat et al., 2014b,a), for which several studies have been conducted to achieve automated diagnosis (Antoniades et al., 2018; Hua et al., 2019; Ansari et al., 2019; Schaper et al., 2019; Acharya et al., 2018b,a), and fMRI. Functional connectivity has shown promise in localizing characteristic differences for autism in resting activity to specific large-scale brain networks (Wang et al., 2018). Whilst there is cautionary evidence using the ABIDE dataset and others (Plitt et al., 2015), it would appear that statistically significant differences in connectivity are generally observable, but like measurements of brain structure, are variable in their presentation. With consistent and localized changes remaining elusive, a number of studies have characterized autism as exhibiting under-connectivity in certain areas of the brain (Just et al., 2004; Cherkassky et al., 2006; Kennedy and Courchesne, 2008; Assaf et al., 2010; Jones et al., 2010; Weng et al., 2010), while others show evidence of over-connectivity (Cerliani et al., 2015; Chien et al., 2015; Delmonte et al., 2013; Di Martino et al., 2011; Nebel et al., 2014a,b). A recent review (Hull et al., 2017) posited that autism is likely a mix of these traits.

Neural networks (LeCun et al., 1999; Hinton et al., 2006; Krizhevsky et al., 2012) are especially adept at classifying complex and large data which parametric inferential statistics may fail to fully characterize due to their inherent assumptions. Given that brain function in autism has been consistently found to be different but in different ways, such a model may be a sensible approach for a comprehensive representation. Previous efforts to classify functional connectivity in autism on smaller datasets have achieved accuracy rates that have been described as "modest to conservatively good" (Hull et al., 2017), though these methods have had trouble replicating on different data (Jung et al., 2014; Price et al., 2014; Iidaka, 2015). More recently, the application of convolutional CNNs to ABIDE data has achieved achieved 68% to 77.3% classification accuracies. (Subbaraju et al., 2017; Brown et al., 2018; Heinsfeld et al., 2018; Khosla et al., 2018).

In this chapter, I leverage the functional connectomes presented in Chapter 2, automatically pre-processing a total of 43,838 functional MRIs from nine different collections. To test the application of CNNs to imaging data, I first classify autistic individuals from typically developing (TD) controls. To validate the proposed models, I then classify functional connectivity matrices based on sex and task vs resting state. All classifications were undertaken using a CNN that uniquely encodes multi-layered connectivity matrices, using an original deep learning architecture, partially inspired by Kawahara et al. (2017). Due to the stochastic properties of NNs and set divisions, I used a standard stratified cross-validation strategy, performing each test across 300 independent models using different subsamples and divisions of the total dataset. To incentivise the model to classify based on phenotypic differences rather than centre differences, a class-balancing technique across participant age and collection were used when building the training and test sets, and compared against the fully-inclusive samples.

4.2 Methods

4.2.1 Datasets and preprocessing

Data were pre-processed using the fMRI Signal Processing Toolbox (SPT), and the full description of extracting functional time series from data is described in Chapter 2. After pre-processing, each dataset was transformed into N $4\times116\times116$ connectivity matrices, using edges weighted by the Pearson correlation of the wavelet coefficients of the pre-processed

					Age				Sex		Disorders
Collection	Subjs	FCs	Rest	Task	Min	Max	Mean	Stddev	F	M	Autism
1000 FC	764	764	764	0	7.88	85.00	25.76	10.18	443	321	0
ABCD	1319	9205	4043	5162	0.42	11.08	10.08	0.65	4339	4866	113
Abide	193	193	193	0	9.00	50.00	17.81	6.69	21	172	94
Abide II	720	761	761	0	5.22	55.00	14.44	7.45	174	587	375
ADNI	141	261	261	0	56.00	95.00	73.57	7.32	146	115	0
BioBank	11811	16970	9937	7033	40.00	70.00	55.23	7.51	8752	8218	8
ICBM	112	381	29	352	19.00	74.00	43.53	14.83	188	193	0
NDAR	1123	8569	5952	2617	0.25	55.83	18.65	7.82	4165	4404	994
Open fMRI	1443	6655	1169	5486	5.89	78.00	27.22	10.40	2768	3133	127
All	17614	43838	23109	20650	0.25	95.00	33.05	20.68	20996	22009	1711

Table 4.1: Average populations present for successfully-preprocessed datasets. Some datasets were not labeled with respect to one or more covariates, so counts may not sum to the listed total.

time-series in each of four frequency scales: 0.1-0.2 Hz, 0.05-0.1 Hz, 0.03-0.05 Hz, and 0.01-0.03 Hz. Because different collections contained fMRIs that used different TRs and sampling rates, wavelet correlation estimates were adjusted to equalize the frequency ranges across different collections. Due to the volume of datasets, individualized quality control was not possible. The proportion of datasets failing pre-processing varied by collection.

Across all collections, 70,284 potential datasets were identified of which 67,396 contained suitable functional and structural datasets. Of these, 52,396 succeeded pre-processing to parcellation. However, datasets with regional dropout of greater than 10% were omitted from the analyses, and redundant datasets across collections were also discarded along with those data with a TR outside of the desired range. In total, 43,838 connectomes from 17,614 unique participants were available for analysis with the NN. Multiple instances of connectomes from the same individuals were used, though they were not shared between the training, validation, and test sets. The numbers of participants, total numbers of datasets used as well as phenotypic distributions, are shown in Table 4.1.

4.2.2 Neural network model and training

The data used for training and testing the CNN were $4 \times 116 \times 116$ (4 wavelet scales and 116 nodes) symmetric functional connectivity (wavelet coefficient correlation) matrices, with

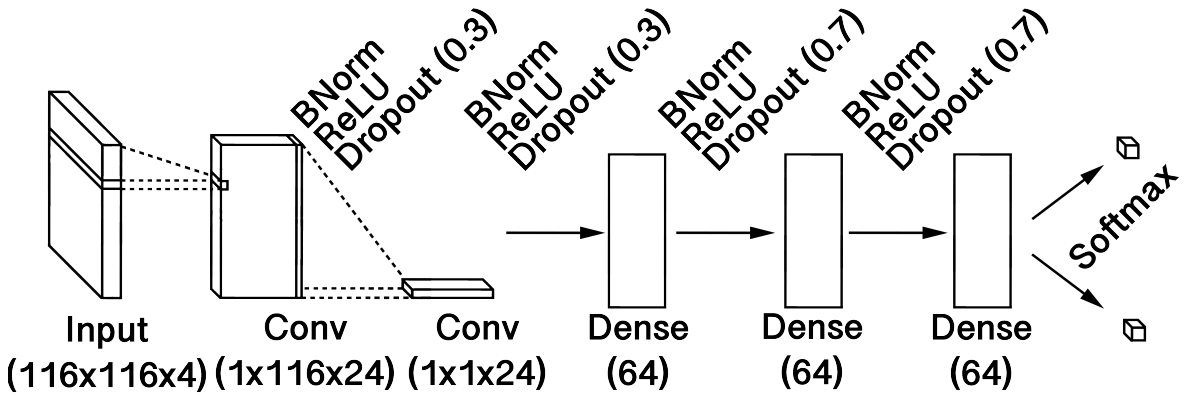


Figure 4.1: The structure of the neural network. These were applied in an ensemble model, so the outputs of 300 independently-trained neural networks were averaged in a cross-validation scheme.

values linearly scaled from [-1,1] to [0,1] for easier use in a NN.

To classify the data, I employed a CNN with vertical convolutional filters on the first layer followed by horizontal convolutional filters on the second layer, effectively reducing the matrices to single values to allow the network to train on connectivity matrices (Figure 4.1). This approach was partially inspired by the cross-shaped filters described in Kawahara et al 2017 (Kawahara et al., 2017), though previous tests with that architecture resulted in a number of failed models with no apparent increase in accuracy over the simpler architecture proposed here.

The CNN was constructed with: 24 edge-to-node vertical convolutional filters; 24 node-to-graph horizontal convolutional filters; 3 fully-connected layers, each with 64 nodes; and a final softmax layer. Separating each layer were batch normalization, rectified linear unit (ReLU), and dropout layers, with the dropout being 0.3 in the convolutional layers and 0.7 in the dense layers. The layer structures and ordering followed the advice offered in Ioffe and Szegedy (2015). Specifications are shown in Figure 4.1. No pooling layers were used, and all strides were of length 1. The model was trained using an Adam optimizer with batch sizes of 64. Otherwise, Keras defaults were used. Models were trained for 200 epochs, and the epoch with the highest validation accuracy was selected.

To obtain a reliable average, I trained 300 models independently for each classification, which were then combined in an ensemble model. In each training instance, a subset of the total available data was taken. A holdout test and validation set were not used (Kohavi, 1995), but instead a division of the data was performed for each model in a stratified cross-

validation schema, subject to the rules detailed below. The ensemble scheme combined these 300 independent models with their own training/test/validation sets by averaging the predictions of any one datapoint across all test sets it was in. Thus, if functional connectivity matrix A appeared in 50 of the 300 test sets, its final output prediction is the average of the 50 independent scores. Datapoints that appeared in at least one test set are considered as being included in the ensemble, while datasets that did not appear in any test sets are excluded from consideration in the final evaluation of model performance.

4.2.3 Set division

Data were divided into three sets: a training set, comprising two-thirds of the data and used to train the model; a validation set, comprising one-sixth of the data and used to select the epoch at which training stopped; and a test set, used to assess the trained classifier performance, comprising one-sixth of the data. The approximate total number of images used by each model was 10,000 for the sex and resting-state classification, and 4000 (limited by sample size) for the autism classification.

For all classifications, a rudimentary balancing algorithm was used such that each class comprised approximately half of the datasets. To account for covariates, classes were additionally balanced such that the distributions of different collections and ages were equal between classes. For collection balancing, equal numbers of datasets were used from each collections. For continuous age values, distributions of age between classes were made to fail a Mann-Whitney U-test, with $p > 0.05$. Standard stratified cross-validation, rather than a holdout division, was used across the 300 runs.

Because of the collection balancing procedure, many data were excluded from certain classification tasks; for instance, as BioBank only included eight subjects with autism. Due to the class balancing, set divisions were not precise in each instance.

4.2.4 Test set evaluation

Inter-data classification

Following the training of the models, the accuracy and the area under the receiver operating characteristic curve (AUROC) were calculated as measures of machine learning performance

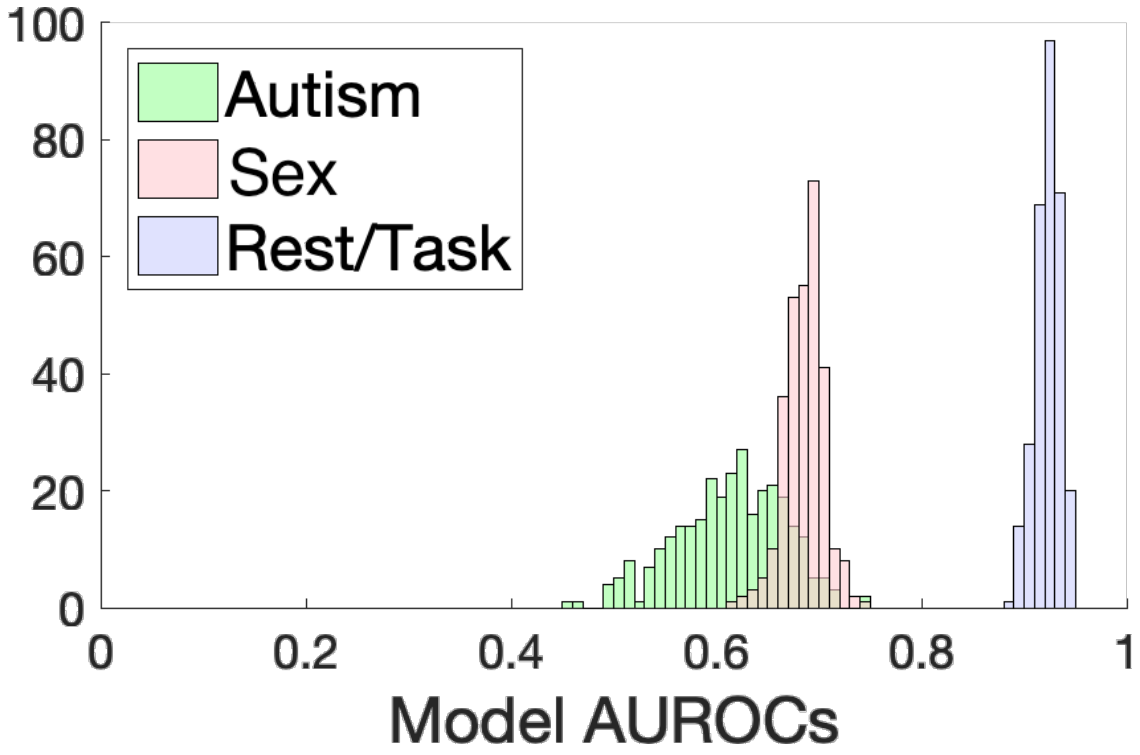


Figure 4.2: Histograms of all AUROCS for 300 independent models, using different, stratified samples of the whole dataset.

on the test set. This was to determine if one group in the classification outperformed the other in training leading to a biasing of the overall accuracy.

Projection of ensemble upper limit

The total accuracy of an ensemble model increases with the number of independent models. Assuming an upper limit to the accuracy that can be achieved by adding more models to the ensemble, I measured the AUROC for random samples of 1 to 300 models and fit this relationship to a logarithmic curve ($y = \frac{a}{1+be^{-kx}}, k > 0$), in which a is the upper limit, predicting the accuracy in the limit of a large number of independent models.

4.2.5 Experiments

I performed the classification on class- and age-balanced datasets that then classified based on sex, task vs rest, and autism vs TD controls in separate analyses.

	Autism	Sex	Rest v Task
Ensemble AUROC	0.6774	0.7680	0.9222
Ensemble Acc.	67.03%	69.71%	85.20%
Average AUROC	0.6133	0.6858	0.9231
Average Acc.	57.12%	63.34%	84.32%

Table 4.2: The ensemble and averaged AUROCS and accuracies for 300 models.

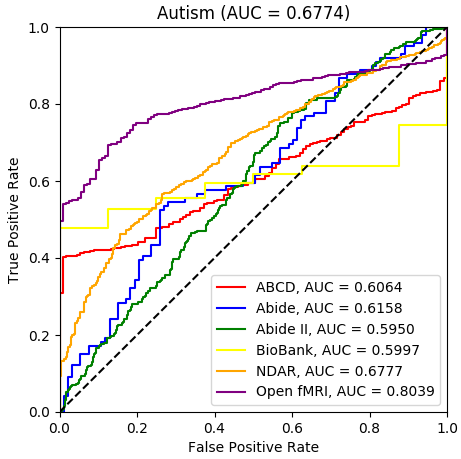


Figure 4.3: The overall classification AUROC and the AUROC of individual data collections for autism classification, showing the overall and relative success of the model.

4.3 Results

Table 4.2 shows the accuracies for the 300 models tested. The AUROCs for the individual models, across all data (Figure 4.2) were averaged to give 0.6858, 0.9231, and 0.6133 for sex, task vs rest, and autism vs TD classifications, respectively, while the average accuracies were 63.33%, 84.31%, and 57.11%. In nearly all cases, however, as shown in Table 4.2, the ensemble AUROC and accuracies were substantially higher. The ROC of ensemble models with respect to collections are shown in Figures 4.3, 4.4, and 4.5.

4.3.1 Autism vs TD controls

With class balancing, the ensemble performance for autism v TD controls across test sets was AUROC=0.6774 (Figure 4.3). Autism classifications were highly dependent on the collection

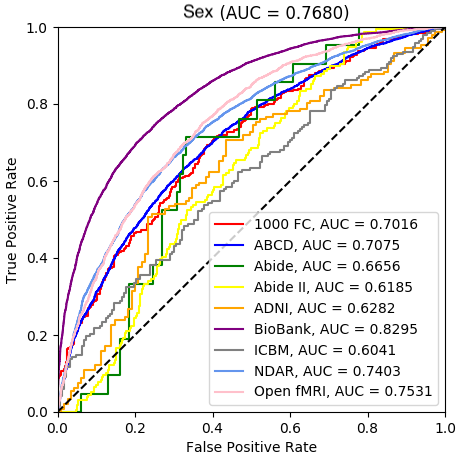


Figure 4.4: The overall classification AUROC and the AUROC of individual data collections for sex classification, showing the overall and relative success of the model.

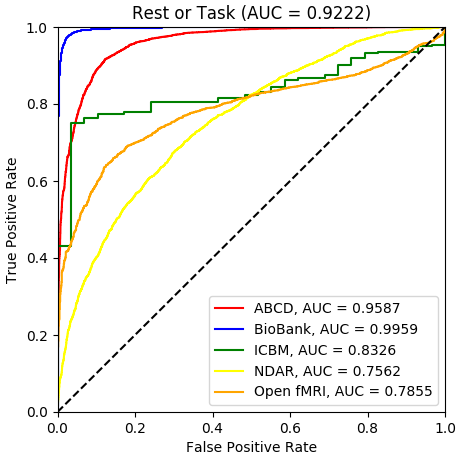


Figure 4.5: The overall classification AUROC and the AUROC of individual data collections for resting-state/task classification, showing the overall and relative success of the model.

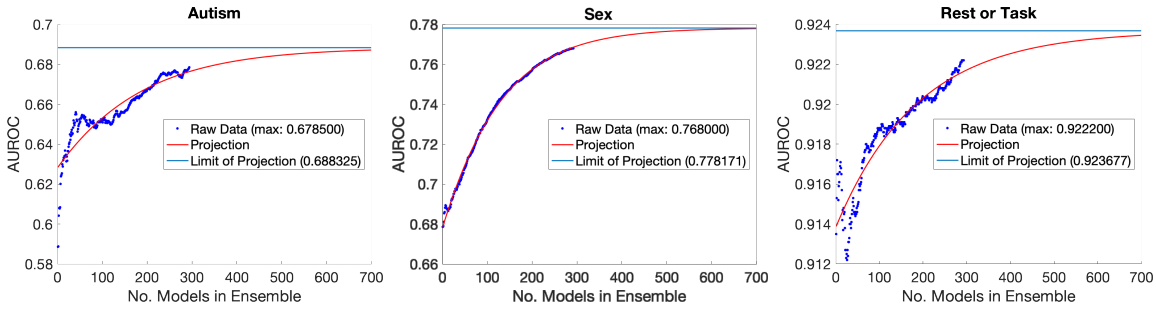


Figure 4.6: Projection of the model limits for AUROCs vertical-filtered CNNs in autism, sex, and resting-state/task classification tasks. The raw data is plotted, as well as the projection of this trend using a logistics growth model ($y = \frac{a}{1+be^{-kx}}, k > 0$), which assumes a hard upper limit (a) to the classification accuracy that can be achieved by simply increasing the number of models in the ensemble. The model predicts that simply adding more models to the ensemble beyond 300 achieves limited returns. In each case, the first ten datapoints were excluded from the model fitting.

used, although the final AUROCs were above chance for all collections. Class balancing was particularly necessary for this scheme, as data from autistic individuals comprised less than 10% overall.

4.3.2 Sex

The ensemble classification of sex yielded 0.7680 AUROC, with comparable AUROCs across different collections (Figure 4.4).

4.3.3 Rest vs task

Task v rest classification had an ensemble classification of AUROC=0.9222 (Figure 4.5), by far the highest of any classification task. BioBank rest/task classification had nearly perfect classification, while other collections that contributed substantial amounts to both resting-state and task participants, that is, NDAR, ABCD, and Open fMRI, had comparable performance.

4.3.4 Ensemble model limits

While the addition of independent CNNs to the ensemble model increases accuracy, this does have a limit. Figure 4.6 shows the plotted AUROC of autism, sex, and rest/task classification. Assuming a hard limit to classification AUROC that can be achieved by simply adding more models to the ensemble, I fitted the data for one to 300 models used in the ensemble to a logistics growth model. This trend showed that, beyond the number of models I have used, there were diminishing returns: in my sex example in Figure 7.4, using another 400 models would lead to a projected increase in AUROC of $0.776 - 0.768 = 0.008$ from 300. However, as shown above, the benefit of using 300 has led to a clear increase in AUROC from just one.

4.4 Discussion

This work describes how large and diverse imaging data might be analyzed by deep learning models, encouraging the aggregation of publicly available collections. Data were partitioned based on clear and logical features of the images and, even with imperfect classification accuracies, deep learning models were capable of recognizing complex patterns in large datasets, many consistent with previous work.

The neuroscientific objective of this study was to use available imaging data with deep learning to describe the pattern of functional brain changes that distinguishes autism from TD. With the absence of any gold standard in this cross-sectional comparison, I also undertook classifications of sex and rest v task, which have more secure, robust findings in the extant literature to confirm the veracity of the developed methods.

In autism, model accuracy was lower compared to the highest rates reported in literature (Brown et al., 2018; Heinsfeld et al., 2018; Khosla et al., 2018), although this result should be viewed with several caveats. The dataset used in this analysis was larger and more varied than any previously analyzed, consisting of many collections. Direct comparisons of machine learning classification methods are difficult as there are no universally accepted schema to divide collections into training and test sets (unlike standardized competitions in other fields, such as the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) (Russakovsky et al., 2015)). Furthermore, my exclusion criteria differed, and, because I opted to use multiple scanning sessions from single subjects during training, I also used follow-up

data in ABIDE not employed in previous studies. Class balancing may also have significantly affected the classification accuracy. However, this was necessary to avoid spuriously large accuracies due to the highly skewed ratios of autism-to-TD individuals. Lastly, preprocessing methods and exclusion criteria are not typically shared across collections, and thus technical and demographic differences in the input data cannot be discounted.

While in this study (and all previous large sample-size studies of autism classification), the classification percentage of autism v TD datasets does not approach the standards of clinical diagnosis, but remains pertinent. First, the intention of the models is to encourage further research and analysis in this field. Second, functional connectivity data may simply lack discrete, distinguishing signals indicative of autism, making perfect classification impossible, in which case deep learning ought to be viewed as an advanced statistical model rather than a potential diagnostic tool. Third, autism is a spectrum and not binary (unlike resting-state/task and, in the vast majority of cases, biological sex), and these labels were applied with varying diagnostic standards. While I am simply using the information available, I recognise that the problem itself may be ill-formed. This is also a potential explanation for the variance in model accuracies seen in Figure 4.2, compared to the other classification problems addressed. Fourth, due to the influence of confounding factors, high accuracy in machine learning for scientific applications should be viewed with skepticism (Ribeiro et al., 2016); for instance, I used several stringent motion-regression algorithms in preprocessing, which likely mitigated the effects of group differences in motion that has previously been observed between autistic and non-autistic subjects (Cook et al., 2013).

Finally, my deep learning model provides several advantages and unique features. First, it employed multichannel input. Although this has long been the standard in 2D image classification (for instance, RGB images), it has not been utilized before in the classification of connectomes. Theoretically, this provides an advantage since it encodes more information about the underlying time-series. In supplementary tests, multichannel inputs generally increased the accuracy of the model by 2–3% over single-channel Pearson correlation input, though this was not tested extensively. Second, it used vertical filters to encode matrices. In initial versions of this study (Leming and Suckling, 2019), I opted to copy the framework of Kawahara et al. (2017), which used cross-shaped filters, although this was found to not increase accuracy over vertical filters and caused the model to sometimes fail. Vertical filters were found to be more compatible with the frameworks of modern deep learning libraries, even though they sacrificed the theoretical advantage of encoding edge-to-edge connections.

This training scheme found substantial accuracy increases with the use of “ensemble” models

in machine learning (Table 4.2); that is, using many independent NNs to vote on a single datapoint. This idea is not new in machine learning (Opitz and Maclin, 1999; Polikar, 2006; Rokach, 2010), but it is notable because the ensemble showed a substantial increase in AUROC and accuracy over the sum of the individual models, and thus in this context it was an effective method of smoothing out unexpected behaviour in models for potential real-world applications. Additionally, it is an effective way to evaluate the performance of a model across the entirety of a dataset, making a good case for classifying functional connectomes using many independent models rather than one.

I showed that more models in an ensemble leads to higher accuracy, though this has diminishing returns after 300. There are two possible explanations behind this trend, and one or both may be the case. The first is that simply adding more models refines the predictions more and more and makes the ensemble less subject to noise. The second is that, with the class balancing scheme used, more models gradually include more and more non-normative test data in their respective training sets, until all data is used at least several times in the prediction, strengthening the overall deep learning model just by the size of its dataset.

It should be noted that the final AUROC was well below the standard for clinical diagnosis, and the variation of model accuracies across our ensemble was very high, especially in relation to the other two categorical classifications. Thus, the areas observed are unlikely to fully characterise autism. This variation across our very mixed dataset is related to the difficulties of diagnosing autism in different contexts, and a binary label applied a spectrum disorder may make for an ill-formed machine learning problem.

4.5 Conclusion

This investigation was the first to amass an exceedingly large and diverse collection of fMRI data and then apply big data methods. I opted to present three important classification tasks and focus on the one that is both most interesting and least-understood. With careful class-balancing, I show that deep learning models are capable of good-quality classifications across mixed collections detecting differences in brain networks, and functions of localized structures, or connectivities over large areas. While the deep learning model in its present form should not be viewed as a diagnostic tool, it is an example of the apparatus needed to statistically analyse large and publicly accessible volumes of data.

Chapter 5

Activation maximization

In this chapter, as a first step in addressing the black box problem, I demonstrate the applicability of activation maximization (a neural network visualization technique) for measuring the relative clustering of different covariates in the deep learning ensemble presented in Chapter 4. By analyzing maximal activations of the hidden layers, I am able to explore how the model organizes a large and mixed-centre dataset, finding that it dedicates specific areas of its hidden layers to processing different covariates of data (depending on the independent variable analyzed), and other areas to mix data from different sources. I present a means of analyzing activation maximization in a single model, then introduce a metric that generalizes this to be applicable to multiple models in the ensemble. This shows that, in spite of collection-, age-, and class-balancing, the models from Chapter 4 nonetheless focused on a number of undesirable confounding factors in its classification.

5.1 Introduction

Deep learning models for MRI classification face two recurring problems: they are typically limited by low sample size, and are abstracted by their own complexity (the "black box problem"). The first of these problems has effectively been addressed by the very large dataset collected, described in Chapter 2. In an initial attempt to address the black box problem, I analyze maximal activations of the hidden layers, allowing for an analysis of how the deep learning model organizes a large and mixed-centre dataset, finding that it dedicates specific areas of its hidden layers to processing different covariates of data (depending on the

independent variable analyzed), and other areas to mix data from different sources. Activation maximization (Erhan et al., 2009) of a hidden layer visualizes how a model partitions a dataset as a whole following classification. I suggest an index to quantify the output of activation maximisation across the ensemble of models.

5.2 Methods

Activation maximization

Activation maximization (Erhan et al., 2009) is a technique to determine the maximally activated hidden units in response to the test set of the CNN layers following training. Activation maximization was applied to the 116×24 second layer of the network (Figure 4.1) as this convolutional layer acts as a bottleneck, and is thus easier to interpret and visualize. This layer is naturally stratified by 24 *filters*, each with 116 nodes (i.e., brain regions in the AAL parcellation). To offset the influence of spurious maximizations, I opted to record the 10 datapoints that maximally activated each hidden unit, obtaining their mode with respect to collection, sex, and whether it was task/rest; for example, if six connectomes that maximally activated a unit were from Collection A and four were from Collection B, Collection A would be recorded as maximally activating that hidden unit.

For each covariate, this method yields a 116×24 array of values for each of the 3×300 models. I opted to measure the stratification of the different convolutional filters in the models by measuring whether it was maximally activated primarily by one source of data, or whether it was activated by a mixed population. With this in mind, I calculated for each layer a diversity coefficient, which is 0 if the layer is only maximally activated by one class of data and 1 if it is maximized proportional to the population maximized. Given K possible classes, $F_k, k \in K$ indicating the proportion of each class in a given filter, and $T_k, k \in K$ indicating the percentage of each class across all filters, I calculated the diversity coefficient for each filter as:

$$D_i = \frac{\tan^{-1}\left(\ln \frac{1-\sqrt{\sum_{k=1}^{K} F_k^2}}{\sqrt{\sum_{k=1}^{K} \frac{(F_k - T_k)^2}{2}}}\right) + \frac{\pi}{2}}{\pi} \tag{5.1}$$

Briefly, the justification for this equation is that the summation $\sum_{k=1}^{K} \frac{(F_k - T_k)^2}{2}$ equals 0 if the

distribution of the filter's population is equal to the population of the whole layer; that is, the distribution is ideally diverse, and this pulls the logarithm towards $-\infty$, which in turn pulls the inverse tangent function to $\frac{\pi}{2}$. Conversely, $1 - \sqrt{\sum_{k=1}^{K} F_k^2}$ tends towards 0 if the individual layer is only composed of a single class, pulling the inverse tangent towards -$\frac{\pi}{2}$. The diversity coefficient is normalized to be between 0 and 1. Its value is indeterminate if only a single class is present globally.

This equation is a more complex version of other diversity coefficients, such as the Herfindahl-Hirschman or Simpson diversity indices. However, the proposed index better accounts for overall populations in the hidden layer activations and thus makes it easier to compare across different classification tasks and independent variables. While the Herfindahl-Hirschman or Simpson indices both approach their maxima when the measured population is completely homogenous, their lower extrema varies depending on the number of distinct populations present. This is problematic in comparing across indices, because the number of populations varies depending on the application, and assumes that the expected (i.e., most diverse) distribution occurs when different populations are perfectly proportional. The proposed index defines the most diverse population as that which has distributions proportional to the overall population, at which point the index is zero.

In practice, low diversity coefficients indicate that the ensemble models stratified data by the covariate. This allows us to measure the degree to which individual covariates (such as collection) were taken into account by the CNNs. I found the diversity coefficient of each of the 24 filters of the hidden, 116×24, convolutional layers, then sorted these values to show which filters were primarily activated by a few covariates and which were activated maximally by many covariates.

5.3 Results

The results in Figures 5.1, 5.2, and 5.3 display the histogram of diversity indices across all models' activation maximisation values. This indicates the tendency of models to use particular filters to sequester data by different covariates, especially if it were attempting to classify by that variable; thus, a diversity index of 0 indicates that all nodes within a particular filter were maximally activated from one or a small number of collections (i.e., BioBank or Open fMRI). The covariates measured are sex, rest/task, and collection site; autism was not included as a covariate because of the relatively small percentage of autism

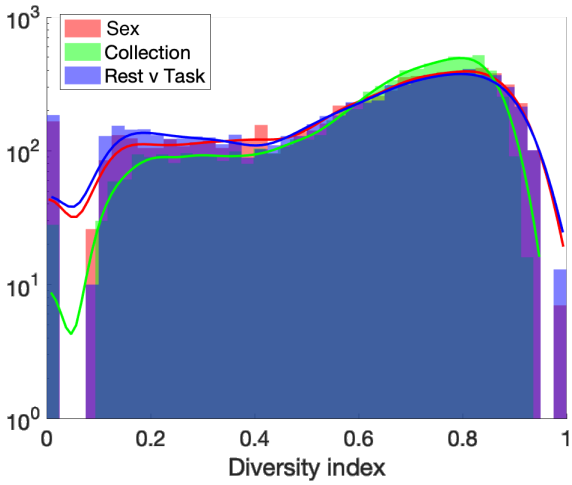


Figure 5.1: The distribution of the diversity index of maximal activations across all filters over 300 models for autism classification, showing how much filters in general were dedicated to particular phenotypes.

data overall.

The diversity index of the activation maximization of the second hidden layer revealed that filters, in many cases, sorted into two distinct groups, as shown by peaks on the lower and upper end of histograms in Figures 5.1, 5.2, and 5.3: stratified layers (i.e., with a diversity index close to 0), which were wholly maximally activated by one type of dataset, and mixed layers (i.e., with a diversity index close to 1), which integrated data from different sources. While sex and task vs rest each had a proportion of their filters wholly activated by a single collection, the majority of filters were activated by a variety of different collections, indicating the effective synthesis of data from different sources. Autism, however, had a large proportion of data with a diversity index close to zero; this is expected for the sex and resting-state covariates, given that the datasets were mainly from males, but the low diversity indices for collection indicates that autism classification models sequestered data based on collection, and thus many datapoints were considered independently.

5.3.1 Autism vs TD controls

Autism classifications were highly dependent on the collection used, although the final AUROCs were above chance for all collections. Activation maximization saw high stratification with regards to sex and resting-state (Figure 5.1). Collection also saw a mix of filters that were both highly stratified and highly diverse, indicating the dual use of convolutional filters.

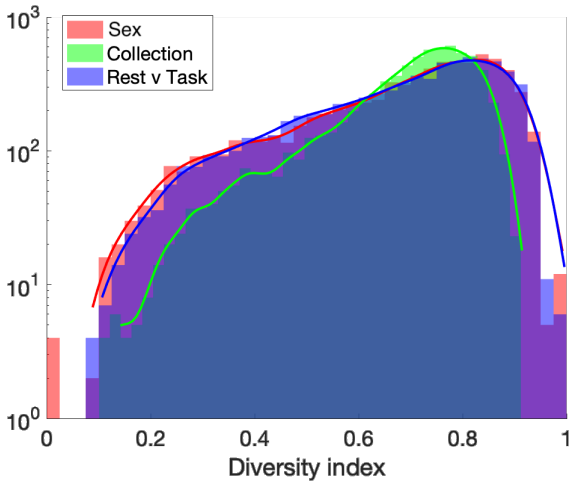


Figure 5.2: This is the distribution of the diversity index of maximal activations across all filters over 300 models for sex classification, showing how much filters in general were dedicated to particular phenotypes.

Given the phenotypic differences in the autism datasets (with ABCD consisting largely of children and ABIDE adolescents, for instance), it is likely that the models considered parts of them independently during classification.

5.3.2 Sex

In activation maximization (Figure 5.2), most of the filters mixed data from different sexes and rest/task. A proportion were maximally activated by individual collections, but for the most part, this was mixed as well. Among the three classification tasks in this study, sex integrated the most data from different sources. As sex distributions are likely the most homogenous variable tracked across datasets (with the exception of ABIDE I and II), the stratification with respect to individual collections was appropriately lower than expected when classifying other variables.

5.3.3 Rest vs task

In activation maximization (Figure 5.3), stratification was found with respect to task (the target covariate), somewhat on collection, and very little with respect to sex. A degree of collection stratification may be expected due to the different tasks found in different collections; for instance, BioBank consisted almost entirely of an emotional faces recognition

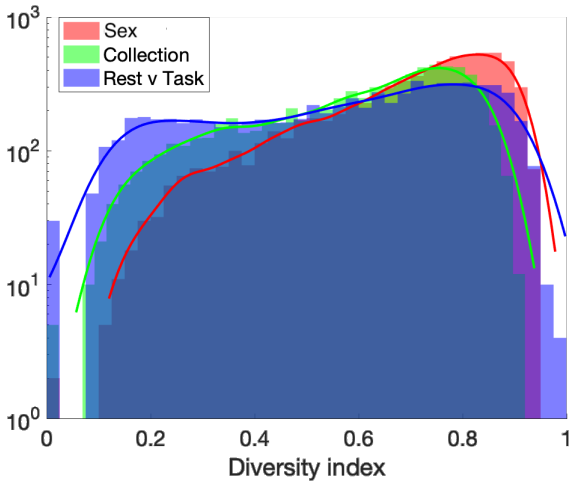


Figure 5.3: The distribution of the diversity index of maximal activations across all filters over 300 models for resting-state/task classification, showing how much filters in general were dedicated to particular phenotypes.

task, while Open fMRI contains a medley of different tasks.

5.4 Discussion

Previously, activation maximization has been used for intuiting the internal configuration of NNs rather than for quantitative interpretation (Erhan et al., 2009), which has never been tried, especially across many different independent models. Many of the filters in these models were wholly activated by datasets from a single group, while others utilized a mixture of datasets. I sought to quantify this effect through a diversity index leading to two general observations: first, across models, a few filters were entirely activated by a single collection (i.e., had a diversity index of 0), though which collection remained inconsistent, and was not apparently proportional to the amount of data contributed by that particular dataset; Sscond, across models, the diversity index was not normally distributed but often had two peaks, one at the low end of the spectrum (indicating stratification of the filters) and one at the high end (indicating a highly diverse, or close to random, distribution of the filter). In autism, a disproportionately high number of filters were activated by a single collection, indicating that the NN split data internally more than other classification tasks.

This chapter shed light on a more general problem with whole-brain MRI classification: even with basic collection and age balancing, the model may still take into account confounding

factors. In the next chapter, this problem is discussed more in-depth, and sophisticated class-balancing techniques are proposed.

Chapter 6

Multivariate class balancing

In Chapter 4, I balanced data by collection and age groups, though, as shown by activation maximization in Chapter 5, there was still a degree of clustering of data during the classification process by different covariates. Eliminating these effects entirely is nontrivial; to mitigate them, however, I introduce in this chapter a more sophisticated class balancing algorithm than that used previously, which can be utilized to regress both discrete and continuous confounding variables. In addition to age, I use this to regress two further confounding factors typically found in MRI data: head motion and intracranial volume. I demonstrate the use of this algorithm for balancing data between different divisions in the data, ensuring that the machine learning model is not incentivized to use confounding factors during classification. This multivariate class balancing scheme ensures equal distributions of these factors within statistical significance.

For this and the subsequent chapter, I choose to focus on a subset of data for one classification task: the UK BioBank, specifically for sex classification. The UK Biobank included both resting-state and task data from a faces/shapes "emotion" task (Hariri et al., 2002; Barch et al., 2013). Details of the acquisition parameters for BioBank data are given elsewhere (Ritchie et al., 2018). After pre-processing, the dataset consisted of 16,970 fMRI acquisitions, decomposed into multi-wavelet-frequency functional connectivity matrices (Patel et al., 2014; Patel and Bullmore, 2016).

6.1 Introduction

Class balancing (often referred to as "dataset matching") is the matching of data from a test group to a control group across a number of discrete or continuous covariates, finding which datapoints between groups are "closest" to one another (in the case of continuous covariates) or which are in the same multiple categories (in the case of discrete covariates). Dataset matching to remove bias from observational studies has been in practice since at least the 1940s (Greenwood, 1945; Chapin, 1947), with a theoretical basis being developed in the 1970s (Cochran and Rubin, 1973; Rubin, 1973). Given the general applicability and need for this practice, development of such methods has been spread across different fields (Stuart, 2010) such as statistics (Rosenbaum, 1989), sociology (Morgan and Harding, 2006), epidemiology (Brookhart et al., 2006), economics (Imbens, 2004), and political science (Ho et al., 2007).

However, the focus of such methods has largely been on small sets of data (Scotina and Gutman, 2019) to simulate randomized control trials for inferential statistics. This field of work is relatively undeveloped in the context of big data for machine learning, for which many computational methods of matching data with continuous covariates would either be computationally intensive, leave out too much, or have not considered the need to find a matching subset of a larger dataset as much as finding test/control divisions. Given the advent of extremely large datasets (Wu et al., 2019), as well as the differences in needs between classical statistics and machine learning methods that use such large datasets (Bzdok et al., 2018), there has emerged a need for alternative methods of dataset matching. With fields ranging from healthcare to economics each having data released by many scattered research groups, matching to synthesize disparate datasets has garnered even more interest (Leulescu and Agafitei, 2013). Recent proposals have even used advanced deep learning models solely to perform this class balancing (Kallus, 2018).

Various class balancing and regression algorithms have covered a large number of different use cases. This chapter proposes a method that is uniquely suited for the purposes of this work, namely balancing a mix of continuous and discrete covariates in a very large (40,000+) dataset that can be measured, but not trivially regressed from, the data itself.

6.2 Methods

6.2.1 Data pre-processing

Pre-processing was completed with the fMRI Signal Processing Toolbox (Patel et al., 2014; Patel and Bullmore, 2016). Following initial identification of the brain parenchyma, and affine registration of the 4D sequence to the mean of the sequence, head motion correction was accomplished using SpeedyPP version 2.0. This process utilized AFNI tools and wavelet despiking (Patel et al., 2014; Patel and Bullmore, 2016), with low- and high-bandpass filters of 0.01Hz and 0.1Hz, respectively, in addition to motion and motion derivative regression. Three motion indicators measured with tools in FSL (FSL motion outliers and FAST; fsl.fmrib.ox.ac.uk/fsl) were recorded that were later applied in class balancing: framewise displacement, spike percentage values (Patel et al., 2014; Patel and Bullmore, 2016), and DVARs (D is refers to temporal derivative of time courses and VARS to root-mean-square of the variance over voxels (Smyser et al., 2010)). Thus, even if motion correction were imperfect, each dataset would have the same distribution of motion values in either class.

Time-series at each voxel in the brain were wavelet despiked to remove transient signals, and then functional and structural datasets were registered to MNI space and parcellated using the 116-area automated anatomical labeling (AAL) template, including subcortical regions (Tzourio-Mazoyer et al., 2002), that defined the nodes of the graph.

The average BOLD signal from each parcel was decomposed by wavelet transform in to three frequency bands: 0.05-0.1 Hz, 0.03-0.05 Hz, and 0.01-0.03 Hz. In each frequency band, separately for each dataset, the correlation of the wavelet coefficients between parcels estimated the edge weights resulting in $N(numberofdatasets) \times 3(waveletfrequencybands) \times 116(parcels) \times 116(parcels)$ symmetric connectivity matrices.

Intracranial volume was estimated from structural images with FSL FAST. This provided an out-of-the-box means of estimating brain volume by deriving tissue types from an input image of the brain. Intracranial volume was estimated by binarizing the GM/WM outputs of FAST, counting the voxels composing this area, and multiplying this by the dimensions of each voxel as read in the NIFTI file header.

This pre-processing was accomplished on a server cluster over a period of several weeks. Due to the volume of datasets, individualized quality control was not possible. From beginning to end, 34.8% of datasets failed the parcellation/wavelet correlation stages and were rejected

from further analysis. Nonetheless, this resulted in 16,970 usable datasets.

6.2.2 Multivariate class balancing

When viewed across the full dataset, there were clear differences in the distributions of covariates when stratifying data by both sex and resting-state/task. Sex differences in intracranial volume are well-documented (Ruigrok et al., 2014), and differences in head motion in resting-state and task datasets were also observed. To address these confounding factors, I implemented an algorithm to balance the datasets such that confounding factors, if successfully measured, were not statistically different between groups. This algorithm first required continuous covariates (such as mean framewise displacement, intracranial volume, and age) to be discretized such that values within a given range are placed into "bins", with each bin covering an equal span of values. Covariates such as collection were already discrete.

6.2.3 Formalization of multivariate class balancing problem

Put in purely mathematical terms, if S is a nonparametric, two-sample test for statistical significance, such that $S(A, B) = 1$ if the null model can be rejected with statistical significance and $S(A, B) = 0$ if it cannot; $A = a_1, a_2...$ and $B = b_1, b_2...$ are datasets A and B, $a_i^{(j)}, 0 < j < J$ and $b_i^{(j)}, 0 < j < J$ one of J (continuous or discrete) measurable confounding factors of the datapoints, then the class balancing problem seeks to optimize the following:

$$\text{argmax}_{|A'|}(A' \subset A, B' \subset B \mid |A'| = |B'| \wedge \sum_{j \in J} S(A'^{(j)}, B'^{(j)}) = 0 \wedge \forall a \in A' \exists! b \in B' : \sum_{j \in J} |a^{(j)} - b^{(j)}| \approx 0) \quad (6.1)$$

This maximizes the number of datapoints included (which benefits the machine learning model) while leaving confounding factors indistinguishable between classes A and B. This has the natural implication of reducing the size of the acquired subsets A' and B' the more confounding variables are regressed.

Notably, the last part of the statement ($\forall a \in A' \exists! b \in B' : \sum_{j \in J} |a^{(j)} - b^{(j)}| \approx 0$) is included to indicate that distributions of confounding variables ought not to be approached independently, and to avoid this, each element in class A' requires an element in class B' that approximately matches it in terms of each measured confounding factor (if the variables are

discrete, this matching must be exact; if they are continuous, the "approximateness" of the matching is dependent on the statistical test).

This issue is best illustrated by a counterexample. Suppose you are training an algorithm to distinguish between a group of cats and a group of dogs, with the confounding factors being sex and fur color. Because neither of those have anything to do with distinguishing the species, the algorithm should not consider them. Suppose that, in a training set, exactly half of the cats consisted of orange-haired females, while the other half consisted of black-haired males; and that half of the dogs consisted of orange-haired males, while the other half consisted of black-haired females. With a dataset such as this, the distributions of sex and fur color are both exactly the same in both groups, yet an algorithm can still achieve perfect accuracy by only considering the confounding factors. However, such a mishap with the training set can be avoided by assuring that there is a one-to-one matching of datapoints between classes with respect to confounding factors.

A further discussion of this problem, as well as the practical complications in finding a global maximum of $|A'|$, is presented in Chapter 9. However, a practical approach to finding A' and B' is described below.

6.2.4 Balancing algorithm

The algorithm curated a subset of the total dataset such that a datapoint from class A within bins $b_1, b_2, ...b_n$ had a corresponding datapoint within the same multivariate bins from class B that was also within the bins $b_1, b_2, ...b_n$. In effect, and bearing in mind that males have larger average intracranial volumes, females with smaller intracranial volumes and males with larger intracranial volumes were used less often in the training set, while males with smaller intracranial volumes and females with larger intracranial volumes were more likely to be included in a particular sampling. There is a tradeoff between the size of individual bins and the size of the dataset, since larger bins are naturally more inclusive, but allow for more variation in the distribution of covariates. Thus, the minimum number of bins was used such that it would not reject the null hypothesis with a nonparametric Mann-Whitney U-test with $p > 0.10$. This algorithm balanced by age, mean framewise displacement (MFD), intracranial volume (ICV), mean DVARs, and mean spike percentage.

This algorithm was applied twice to the data. The first balanced men and women. This scheme forced a 1:1 ratio between sexes, with distributions of respective covariates main-

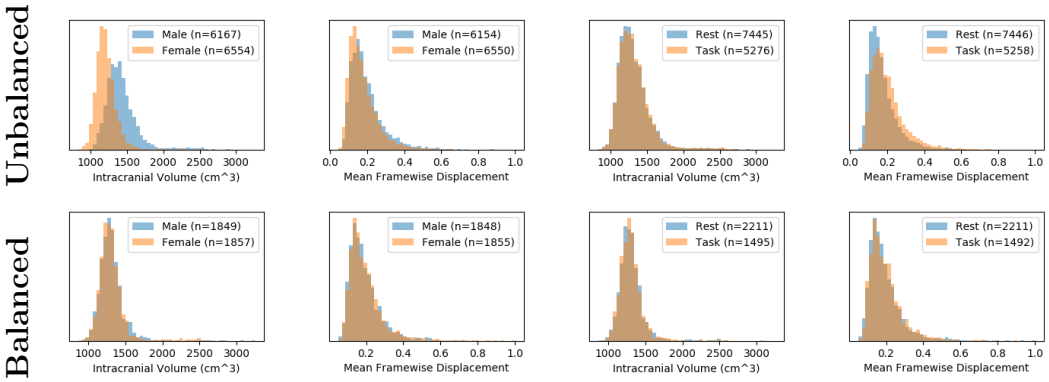


Figure 6.1: Histograms displaying distributions of random training sets with respect to mean FD and intracranial volumes, divided both by gender and resting-state/task, before and after the class balancing scheme.

tained. Data was then balanced by resting-state and task, though no ratios were forced. This left four divisions in the data: resting-state and task, men and women, with approximately equal distributions of confounding factors.

6.3 Results

Prior to class balancing, the datasets displayed significant motion effects between groups, especially with regards to task- and resting-state differences, as well as significant differences in intracranial volumes between sexes (Figure 6.1). The class balancing scheme selectively eliminated datasets such that each class had similar distributions across each covariate, as well as a 1:1 ratio of males to females. The same balancing procedure was also performed for resting-state and task data, with the original ratios present in the dataset maintained. Class balancing disincentivized the model from classifying based on confounding factors. The balanced class distributions can be seen at the bottom of Figure 6.1.

As shown in Figure 6.1, the balanced dataset resulted in a much lower number of usable datasets. This limits the use of this methodology to big data contexts. Nonetheless, when applied in a cross-validation schema with ensemble models, in which a balanced dataset is prepared independently for each model, this still results in the majority of data being used at least once in an ensemble, as will be shown in Chapter 7. However, data that is normative in head motion and intracranial volume is utilized more across all ensembles, while data that has any extrema is naturally excluded.

Chapter 7

Salience in brain connectomes

In this chapter, I introduce a framework for applying deep learning visualization techniques on brain connectome data, thus allowing for the analysis of edge saliency (i.e., which edges contributed the most to the output of classification tasks). First, I briefly demonstrate the use of class activation mapping (Selvaraju et al., 2017) on the results in sex, autism, and rest/task classification from Chapter 4, showing a problem in the encoding techniques of the vertical CNNs from Chapter 4 when edgewise resolution is desired. To solve this issue, I introduce a stochastic encoding method that may be applied in a CNN ensemble to improve resolution. This method was applied to the balanced dataset from Chapter 6, analyzing resting-state and task data from the UK BioBank in sex classification, using class activation mapping and occlusion to measure the salience of three brain networks involved in task- and resting-states, and their interaction. This achieved a final AUROC of 0.8459. The results showed that resting-state data classified more accurately than task data, with the inner salience network playing the most important role of the three networks overall in classification of resting-state data and connections to the central executive network in task data.

7.1 Introduction

While I have shown in Chapter 4 that CNNs can be powerful tools for classifying functional connectomes, they face a problem with interpretability. Even if CNNs can classify data successfully, it is unknown which features of input data make a disproportionate contribution

in the process, and the model remains a black box. Although Chapter 5 presented one method of analyzing the internals of CNNs, it said nothing about features of the input data itself. Knowledge of such features are especially necessary for biological applications in which the underlying mechanisms of the systems being classified are often of the greatest interest. To overcome this issue, a number of ways to visualize and quantify neural networks have been pioneered in recent years. Two such methods include occlusion, in which the classification accuracy is measured when specific input data are systematically omitted from the process (Zeiler and Fergus, 2013); and salience maps (Simonyan et al., 2014), later adapted into class activation maps (Selvaraju et al., 2017), in which the derivative of the neural network with respect to an input datapoint is approximated displaying which parts of the input data effected the most change in the neural network.

This thesis has discussed issues in encoding graphs, particularly brain connectomes, for CNNs. In Chapter 4, I adapted a framework called BrainNetCNN (Kawahara et al., 2017) (which, notably, did use salience maps to analyze data) to use vertical filters to encode based on the column of a connectivity matrix, in order to classify multi-slice functional connectomes. This replaced square-shaped filters that are more typical for 2D image classification.

While encoding based on the columns of a connectivity matrix is intuitively sound, given that it accounts for the edges connected to a particular node, it does in theory have three problems, especially when a salience algorithm is applied. First, the convolutions bias the output class activation maps; a highly salient single edge would also increase the salience of edges in its same row or column. Second, it is difficult to determine the veracity of saliency algorithms from biological data where the ground truth is unknown, as for single runs the algorithms may give spurious results (Kohavi, 1995), whereas they often indicate "visual saliency" for 2D images (i.e., areas of the image on which human subjects focus), which are straightforward to verify by a human observer. Because of the inconsistencies between ML models, the most robust solutions come from averaging salience maps found over a number of trained models (Khosla et al., 2018; Leming and Suckling, 2020a). Third, convolving whole columns or rows with a single value (node) encodes a large amount of input data that scales with the size of the input matrix. This dilutes the relative contributions of single edges which may be essential in classification, and possibly leads to underfitting.

7.1.1 Network brain function across the sexes

Taken on their own, differences found between task-based and resting-state brain activations may be among the most robust discoveries of fMRI studies. The default mode network (DMN) has been consistently identified as a marker of resting-state (i.e. in the absence of a cognitively effortful task) connectomes since it was first described (Raichle et al., 2001). Other brain networks emblematic of particular tasks have been identified as well (Smith et al., 2009), including the dorsal and ventral attention networks (Corbetta and Shulman, 2002; Vossel et al., 2014), which are respectively concerned with voluntary focus on features and switches in attention or unexpected stimuli; i.e., the change between resting-state and task fMRI. As noted by Fox et al. (2005), when performing simple memory tasks, the response commonly observed is proportionally increased activity in certain frontal and parietal cortical regions (Cabeza and Nyberg, 2000; Corbetta and Shulman, 2002) and decreased activity in the posterior cingulate, medial and lateral parietal, and medial prefrontal cortex (Gusnard et al., 2001; Simpson et al., 2001; Shulman et al., 1997; McKiernan et al., 2003; Mazoyer et al., 2001), which form the default mode network. Fox et al. (2005) identified two widely distributed, anticorrelated networks in the brain that exist in the resting state, but intensify during tasks. Additionally, switches between the resting-state and task often involve transitions from the DMN to the central executive (CEN) and salience networks (Goulden et al., 2014). The CEN is the dominant network following suppression of the DMN when a cognitively demanding task is being performed (Fox et al., 2006), while the salience network is activated in a less task-specific manner and more in response to perceived cognitive, homeostatic, or emotional salience (Seeley et al., 2007), which may be brought on by pain, uncertainty, or emotional tasks. Effective connectivity studies with granger causality (Sridharan et al., 2008) and dynamic causal modeling (Goulden et al., 2014) have indicated that the DMN to CEN transition is modulated by the salience network.

Sex differences in brain networks, and more generally the functional processing of tasks, is an area of active scientific interest. But while functional imaging studies of the brain have often found differences between men and women, it is difficult to compare studies due to small sample sizes, differing analysis methods, different areas selected a priori for testing, and differences in particular tasks. Various task fMRI studies have found widely spread sex differences in the bilateral amygdala, hypothalamus, right cerebellum, and posterior and superior temporal sulcus in response to emotional and visuospatial processing (Hamann et al., 2004; Takahashi et al., 2006; Mackiewicz et al., 2006); right hemisphere activation in response to visuospatial tests (Gur et al., 2000); differing activations in the superior parietal

lobule and the inferior frontal cortex in response to mental rotation tasks (Hugdahl et al., 2006); and limbic regions, prefrontal regions, visual cortex, the anterior cingulate gyrus, and the right subcallosal gyrus in response to emotional faces (Fischer et al., 2004; Fusar-Poli et al., 2009).

Three large sample-size neuroimaging studies that documented functional sex differences in resting-state fMRI in both developing (Tomasi and Volkow, 2011b; Gur and Gur, 2016) and adult populations (Ritchie et al., 2018) found higher local functional connectivity in women than in men, higher connectivity in the DMN in women, and lower connectivity in the sensorimotor cortices, though unlike the emotional stimuli studies there were no particularly localized differences in activation between the samples. This was possibly due to the higher variation of resting-state fMRI due to its unconstrained nature (Buckner et al., 2013; Elton and Gao, 2015).

When classifying between sexes, past ML studies using methods ranging from support vector machines to CNNs, have achieved classification accuracies between 65% and 87% (Casanova et al., 2012; Satterthwaite et al., 2015; Gur and Gur, 2016; Zhang et al., 2018), depending on the dataset and methods used. In Chapter 4, I performed a classification by sex of functional connectomes acquired at multiple sites using a CNN with vertical filters, with a final area under the receiver operating characteristic curve (AUROC) of 0.7680, including an AUROC of 0.8295 with single-site, UK BioBank data. Additionally, DTI data classification has led to exceptionally high accuracies (93%) (Anderson et al., 2019; Xin et al., 2019), though such modalities are not always readily available.

The effects of sex on macro resting-state and task networks are still debated (Goldstone et al., 2016). Some studies (Liu et al., 2009; Agcaoglu et al., 2015) have found that sex modulates the lateralization of resting-state networks, while other studies have reported only a small (Bluhm et al., 2008; Lopez-Larson et al., 2011) or non-significant effect (Weissman-Fogel et al., 2010; Nielsen et al., 2013a). Network-level sex differences in task fMRI indicate that men and women process tasks differently. Adolescent females have been reported as having higher functional connectivity in the DMN and fronto-parietal networks during a self-referential processing task (Alarcón et al., 2018). Analysis of canonical networks in task fMRI, although not able to draw substantial conclusions on the roles of the networks in different tasks, found that tasks involving fluid intelligence were the most discriminative for sex (Greene et al., 2018). These studies would suggest that men and women process tasks differently. However, they have not been validated on larger datasets.

Table 7.1: Table of the averaged and ensemble AUROCS of the models run in this paper. Each row represents a batch of 300 independent models. The complete model, which considered all edges, is shown in the top row. The next six rows showing the results of classifying half of the edges including and excluding those edges inside a given network (45 unique edges total). The next six rows show the same results, only considering the edges connecting to those networks as well (1105 edges total).

		All		Rest		Task	
		Ens.	Mean	Ens.	Mean	Ens.	Mean
Complete		0.8459	0.8010	0.8923	0.8504	0.7683	0.7207
Inner Edges Only							
CEN	Incl.	0.8380	0.7805	0.8844	0.8343	0.7609	0.7027
	Excl.	0.8386	0.7798	0.8825	0.8315	0.7641	0.7050
DMN	Incl.	0.8407	0.7804	0.8868	0.8336	0.7643	0.7018
	Excl.	0.8420	0.7806	0.8873	0.8334	0.7671	0.7030
SAL	Incl.	0.8388	0.7824	0.8860	0.8352	0.7600	0.7050
	Excl.	0.8392	0.7782	0.8853	0.8308	0.7631	0.7021
Connecting Edges							
CEN	Incl.	0.8406	0.7833	0.8872	0.8364	0.7624	0.7059
	Excl.	0.8287	0.7704	0.8738	0.8228	0.7544	0.6939
DMN	Incl.	0.8396	0.7801	0.8836	0.8337	0.7660	0.7020
	Excl.	0.8278	0.7712	0.8753	0.8246	0.7490	0.6929
SAL	Incl.	0.8397	0.7811	0.8875	0.8351	0.7619	0.7024
	Excl.	0.8321	0.7739	0.8853	0.8253	0.7631	0.6993

The objective of this chapter is not only to utilize CNNs to classify functional connectomes, but explain the classification performance in terms of those edges and subnetworks that are most salient. To do so, I introduce a stochastic deep learning model that allows for the consideration of each edge in a network independently without overfitting, presenting robust results by training and combining many such models in the ensemble framework introduced in Chapter 4. Convolutions with random samples of edges allow for the consideration of each edge independently without overfitting, and in training many such models and averaging their outputs, this effectively addresses all of the issues with class activation maps outlined above.

Combining this with the dataset resulting from the multivariate class balancing scheme

introduced in Chapter 6, I characterized sex differences in connectomic representations of resting-state and task fMRI (in UK Biobank data, a faces/shapes "emotion" task (Hariri et al., 2002; Barch et al., 2013)) with a focus on the DMN, the salience network, and the CEN. I evaluate performance with the average AUROC across 300 models in the ensemble scheme. To further justify the use of stochastic encoding, I applied guided gradient class activation mapping (Grad-CAM) to the results from Chapter 4, showing the effects of this on vertical-filtered CNNs. I then used Grad-CAM (Selvaraju et al., 2017) and occlusion (Zeiler and Fergus, 2013) of individual brain networks to evaluate the salience of each edge within and connecting to brain networks for the class-balanced UK BioBank data on stochastic CNNs, comparing their relative salience within the model.

7.2 Methods

7.2.1 Machine learning

I classified functional data from the UK BioBank by sex. Because classification of UK BioBank rest/task data achieved near-perfect accuracy in Chapter 4, I chose not to repeat this analysis. Here, the focus was on the relative classification accuracy of task data and resting-state data when classifying by sex.

Model structure

The deep learning model was an ensemble of stochastic CNNs. The architecture is shown in Figure 7.1. I first randomly permuted the columns (nodes) of the connectivity matrices, preserving the permutation order across wavelet frequency bands. These matrices were then input to a CNN with 256 filters of shape $1 \times 58 \times 1$. This convolved 58×3 random values of the matrix which was then fed into three dense layers, each with 64 hidden units, with batch normalization layers, rectified linear unit (ReLU), and 0.5 dropout between them. Finally, the data was binary classified through a softmax layer.

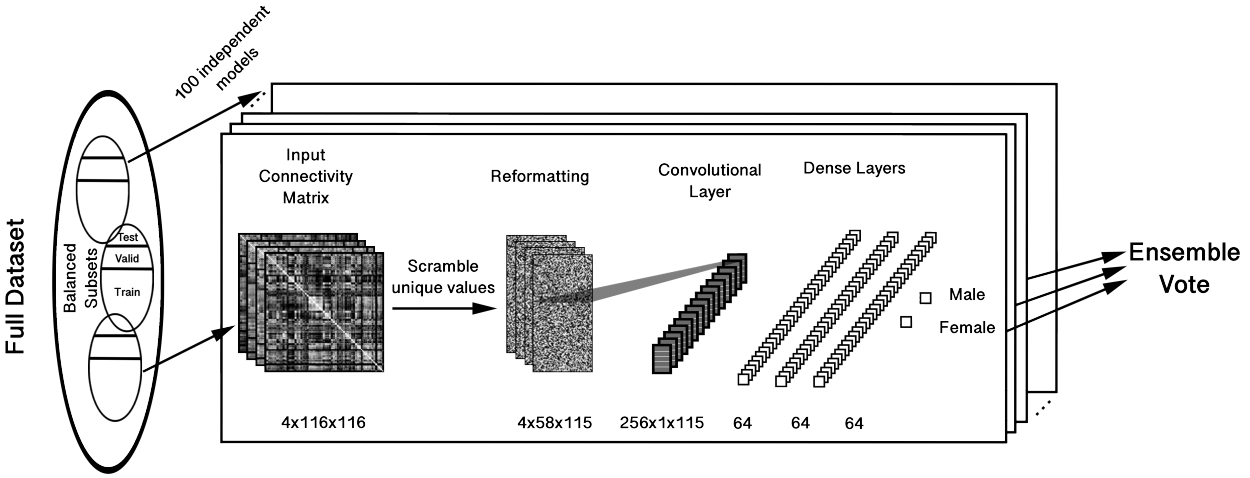


Figure 7.1: In this model, matrices are encoded by random scrambling prior to being fed into a single convolutional layer, followed by three dense layers. In between each layer is a batch normalization and rectified linear unit (ReLU) layer, with 50 percent dropout in between the dense layers. Our training scheme trains 300 such models, each with its unique scrambling order, independently on a class- and covariate-balanced subset of the whole dataset, then combines votes for datapoints appearing in overlapping test sets into a final ensemble vote.

Training

The data were separated into training, validation, and test sets, with an approximate ratio of 4:1:1. I trained 300 CNN models on random class-balanced subsamples of the whole dataset. Each model was trained for 100 epochs (cycles through the training set), and the epoch with the highest validation accuracy was selected. CNN performance was reported on the test set. These 300 models with their respective test set classifications were then unified in an ensemble model. The output classification of a dataset appearing in $\frac{n}{300}$ models was averaged across n models. Thus, datasets were not counted more than once when measuring the final accuracy of the ensemble models, reported as AUROCs. In total, 14,683 datasets were used at least once in the test sets, comprising 86.5% of the overall dataset.

7.2.2 Visualization of machine learning results

I used two different ML visualization methods to assess the role of three different, a priori brain networks in the sex classification of resting-state and task data.

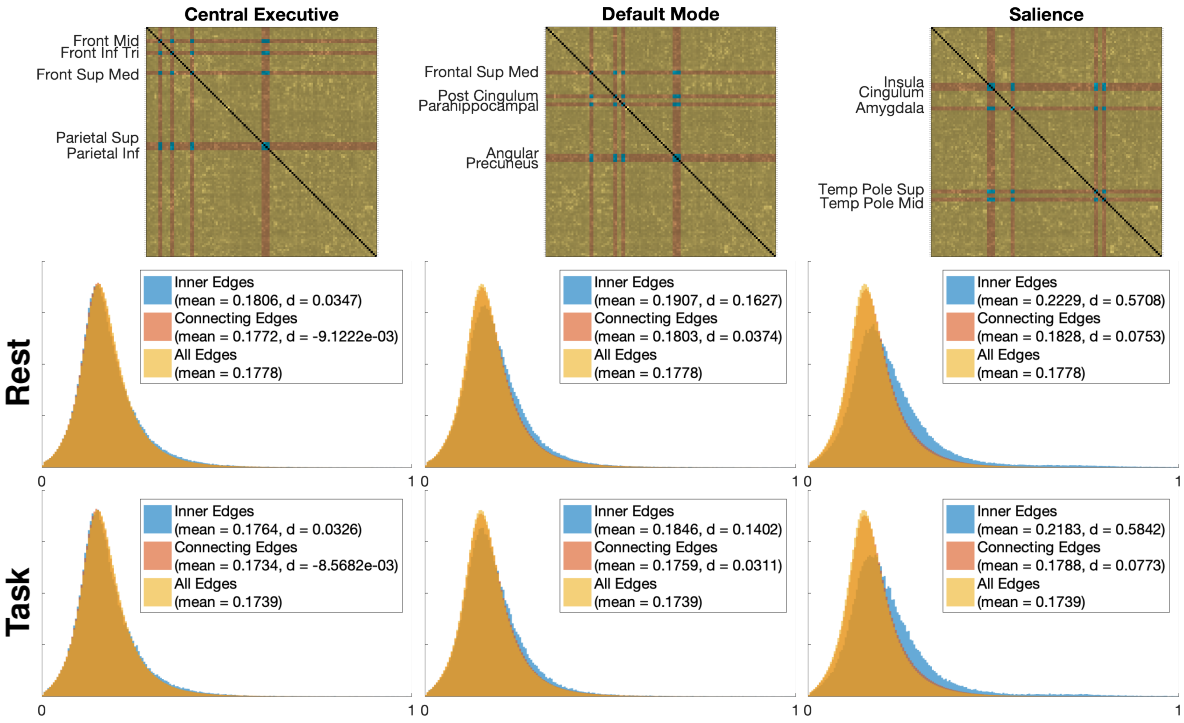


Figure 7.2: (Top) The averaged class activation maps (CAMs) across all subjects for the complete graph classification, with the three studied networks highlighted. Area names in the AAL atlas are given. (Bottom) Histograms of all inner and connecting CAM values of the three networks, both in resting-state and task subjects, compared to the overall distribution of CAM values. Because the large number of samples, we display the effect size (measured by Cohen's d) of both inner and connecting edges compared to the CAM values of the rest of the edges.

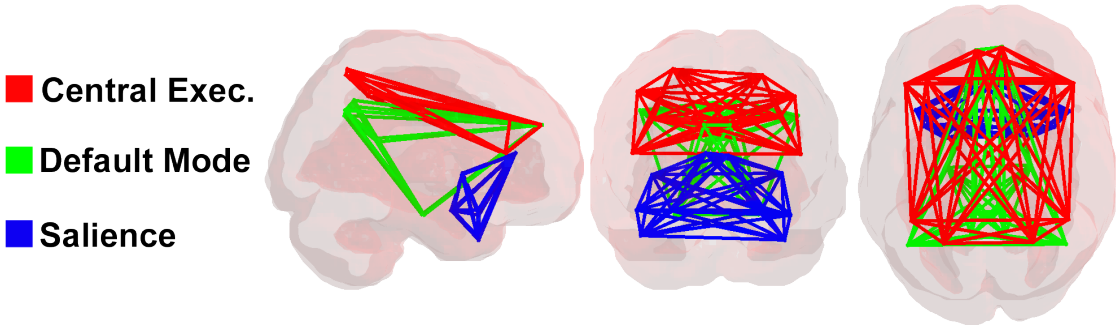


Figure 7.3: A 3d display of the three networks analyzed in this paper, in the AAL parcellation. Green: default mode network; blue: salience network; red: central executive network. Each network is comprised of ten distinct brain regions.

Brain network encoding

To assess the role of the DMN, CEN, and salience network in classification, I selected representative nodes from the AAL parcellation (named in Figure 7.2), referring to prior network descriptions (Mulders et al., 2015). Each network comprised 10 distinct nodes. The DMN was characterized by a combination of the medial frontal gyrus, posterior cingulum, parahippocampus, precuneus, subgenual anterior cingulate cortex, and inferior parietal lobe, the CEN by the bilateral middle frontal lobe, frontal interior triangularis, frontal superior medial, and the superior and inferior parietal lobe, and the salience network by the bilateral insula, anterior cingulum, amygdala, and the middle and superior temporal pole (Figure 7.3).

For both of the analysis methods described below, I isolated edges making up these networks in two different ways: first, by exclusively selecting edges within the network; i.e. edges connecting two nodes of a given network (comprising $\frac{10\times(10-1)}{2} = 45$ unique edges); and second, all edges within, and connecting to a network, by selecting those edges that connect to at least one other node (comprising $10\times(116-1)-\frac{10\times(10-1)}{2} = 1105$ unique edges). Thus, for each analysis method, two sets of results are presented: one for the sets of edges within a network, and the other for all edges connected to a network.

Gradient class activation maps

I applied the Grad-CAM algorithm (Erhan et al., 2009; Selvaraju et al., 2017; Kotikalapudi and contributors, 2017) to find class activation maps (CAMs) for each dataset in each CNN model. Grad-CAM is an extension of the general salience algorithm (Simonyan et al., 2014). In its simplest form, salience is obtained by taking the derivative (approximated as a first-order Taylor expansion) of a particular deep learning model with respect to a particular input image. In studies of 2D images, CAMs are able to distinguish between different objects within a single image belonging to different classes (Selvaraju et al., 2017); for example, in a multiclass classifier of a picture of a cat and a dog, taking an image with respect to class 0 would highlight the cat, while taking the same image with respect to class 1 would highlight the dog. Grad-CAM extends this by making CAMs applicable to a variety of CNNs, including those that use fully-connected deep layers, as used here.

I first show the need for stochastic encoding when applying Grad-CAM. To do so, I refer back to the results for autism, sex, and rest/task classification from Chapter 4, briefly displaying those results. In this case, CAMs were averaged for each output in the dataset with respect to its measured classification accuracy.

I then derived CAMs from each independent stochastic CNN with respect to both class 0 (females) and class 1 (males) across three wavelet bands and averaged these across the 300 models, producing a single 116×116 CAM for each fMRI dataset in the ensemble models. The total distribution for CAM values within and connecting to each particular brain network was then compared to every other CAM value. Due to the extremely large number of values, distributional differences were measured by Cohen's d (effect size), rather than statistical significance.

Occlusion

In separate sex classification models, I occluded half of the edges for each model in the ensemble and trained on the occluded data. This was inspired by photographic image occlusion (Zeiler and Fergus, 2013) which deliberately excludes portions of data and measures relative classification accuracy with the occluded data as a means of detecting salient areas. The importance of the three brain networks to the classification was tested by comparing the average AUROC of 300 models whose occluded edges were the edges making up the particular brain network, and 300 models for which brain networks were not occluded. I trained on each set using the same 300 model/ensemble scheme detailed above (see Figure 7.9, top). The relative accuracies of these independent models, both on the complete dataset and for the resting-state and task fMRI data, were compared to understand the contributions of different networks to sex classification in both resting-state and task fMRI. In particular, I applied a nonparametric statistical test on the two sets of 300 AUROCs including and excluding a particular brain network, then reported the p-value of this test, corrected for multiple comparisons.

I trained, for each of the three networks, 300 models that included the given network and 300 excluding it, each with the two different encoding schemes (i.e. considering the edges only within a network and all edges connected to a network), for each of the three networks (DMN, CEN, and salience network). In total, I trained $2 \times 2 \times 3 \times 300 = 3600$ models for these occlusion tests.

7.3 Results

7.3.1 Machine learning

Model accuracy

I initially classified by sex-balanced datasets with both resting-state and task fMRI. I used 300 independent CNNs that took as input randomly scrambled unique values of the input wavelet correlation matrices (Figure 7.1) in a stratified cross-validation (Kohavi, 1995) scheme. The final results for the 300 models are given in Table 7.1 (top row) with an average AUROC of 0.8010 when assessing the CNNs independently. However, when all 300 models were aggregated into a single classification such that predictions for a particular dataset appearing across multiple independent models were averaged into a single value (Figure 7.1), the AUROC was 0.8459.

The ensemble model also classified sex in resting-state fMRI with an ensemble AUROC of 0.8923 and task fMRI with an AUROC of 0.7683, a difference of 0.1240. Full results are given in Table 7.1.

Projection of ensemble upper limit

The upper predicted limit of AUROC in the limit of a large number of datasets, based on a logarithmic model, is shown in Figure 7.4, and was found to be 0.8477.

7.3.2 Visualization of machine learning results

Grad-CAM results for vertical filters (Chapter 4)

This phenomenon of dilution of CAMs with vertical filters, which justifies the use of stochastic encoding, may be demonstrated visually by showing the results of averaged class activation maps on the output of the vertical-filtered CNNs from Chapter 4, which are shown in Figure 7.5. A brief explanation of the findings in sex, autism, and rest/task classification from Chapter 4 are as follows:

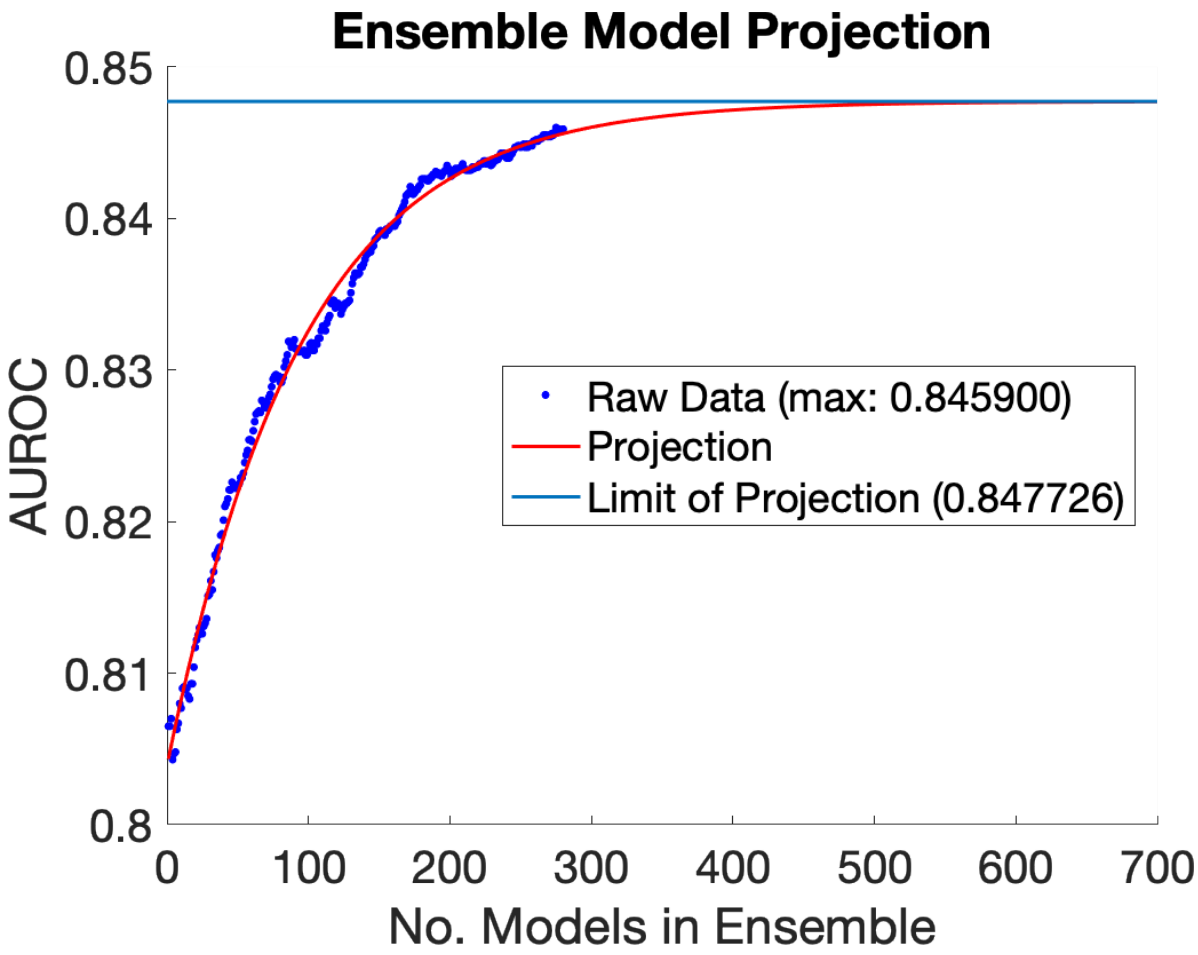


Figure 7.4: Gender classification AUROC across 1 - 300 independent CNNs included in the ensemble model. The raw data is plotted, as well as the projection of this trend using a logistics growth model ($y = \frac{a}{1+be^{-kx}}, k > 0$), which assumes a hard upper limit (a) to the classification accuracy that can be achieved by simply increasing the number of models in the ensemble. The model predicts that simply adding more models to the ensemble beyond 300 achieves limited returns. The upper limit is 0.8477, with 95% confidence bounds between 0.8473 and 0.8481.

Autism vs TD Controls

Class activation was strongest for autism in the limbic system, cerebellum, temporal lobe, and frontal middle orbital lobe, but overwhelmingly emphasized in the right caudate nucleus and paracentral lobule (Figure 7.6). Findings of the caudate nucleus are consistent with historical findings in developmental autism (Qiu et al., 2015), with both aberrant functional connectivity frequently associated with that area and the presence of volume differences (Sears et al., 1999; McAlonan et al., 2002; Brambilla et al., 2003; Hollander et al., 2005;

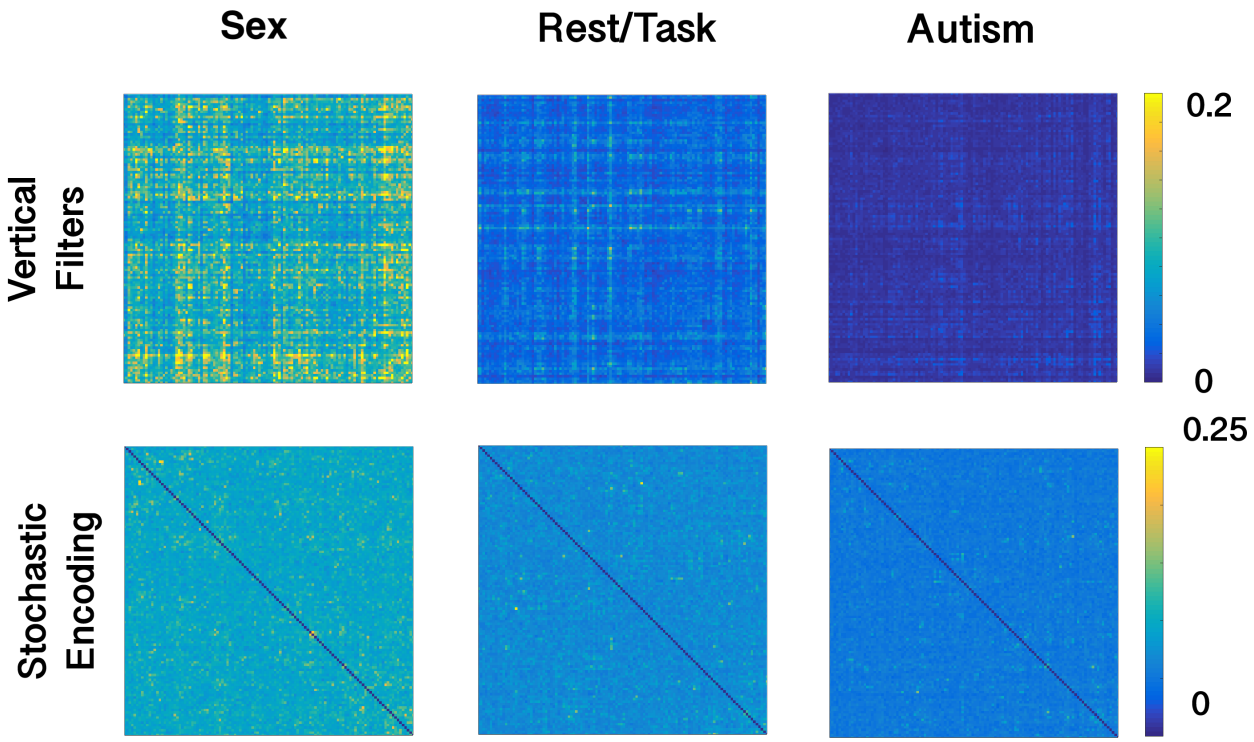


Figure 7.5: Comparison of the encoding classification tasks from Chapter 4 between vertical filters and stochastic encoding, as displayed in averaged class activation maps. This displays visually the bias of vertical filters in the classification, whereas stochastic encoding allowed for finer resolution of the salience of particular edges.

O'Dwyer et al., 2016; Rojas et al., 2006; Turner et al., 2006; Qiu et al., 2016).

Sex

On average, CAMs in sex classifications showed more differences around areas in the corpus callosum and the frontal lobe (especially the medial left frontal lobe), as well as parietal areas, with very few subcortical differences (Figure 7.7). Note that this includes the full dataset and not just the UK BioBank data.

Rest vs task

The CAM (Figure 7.8) focused on the default mode network, largely in the left hemisphere, and its connection to the right frontal medial orbital area. The highly emphasized areas include the supplementary motor area, the left parietal lobe, the bilateral middle and inferior occipital lobe, the left precentral gyrus, and the bilateral thalamus representing the wide range of areas activated in task fMRI.

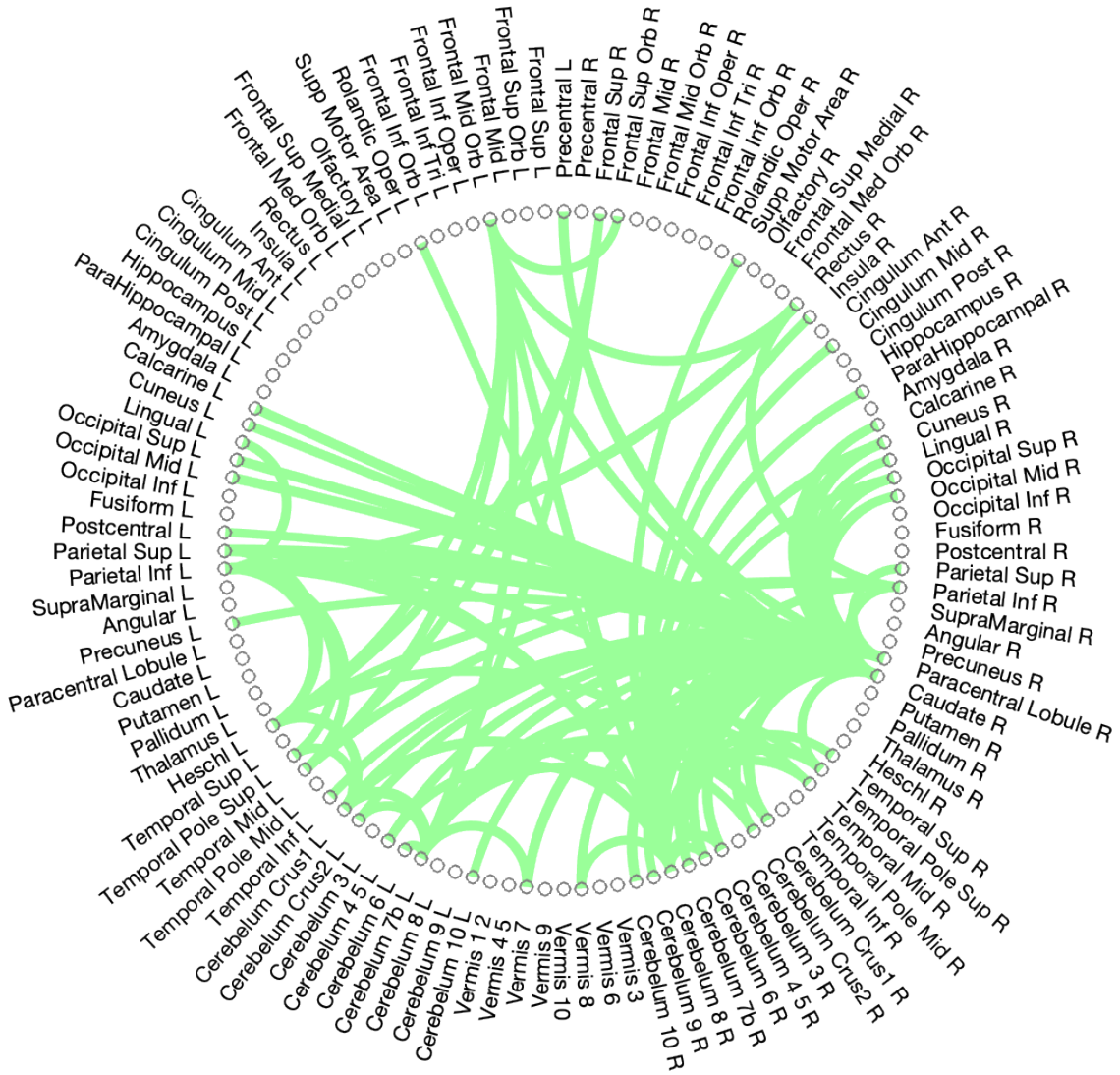


Figure 7.6: The 100 strongest connections of the mean class activation maps in autism classification for the vertical-filter model, with the maximum value taken across wavelet correlations.

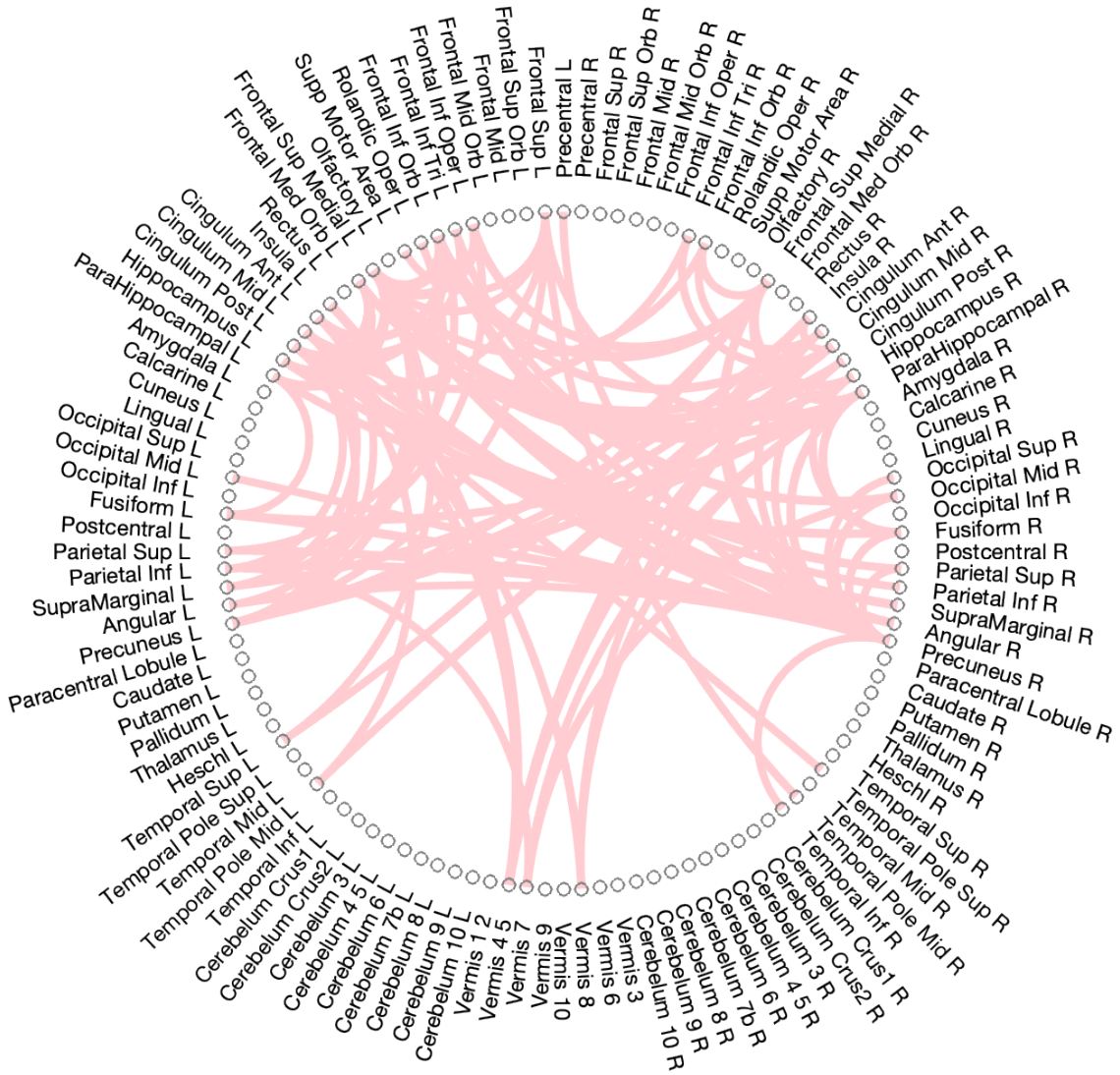


Figure 7.7: A histogram showing the 100 strongest connections of the mean class activation maps in sex classification for the vertical-filter model, with the maximum value taken across wavelet correlations.

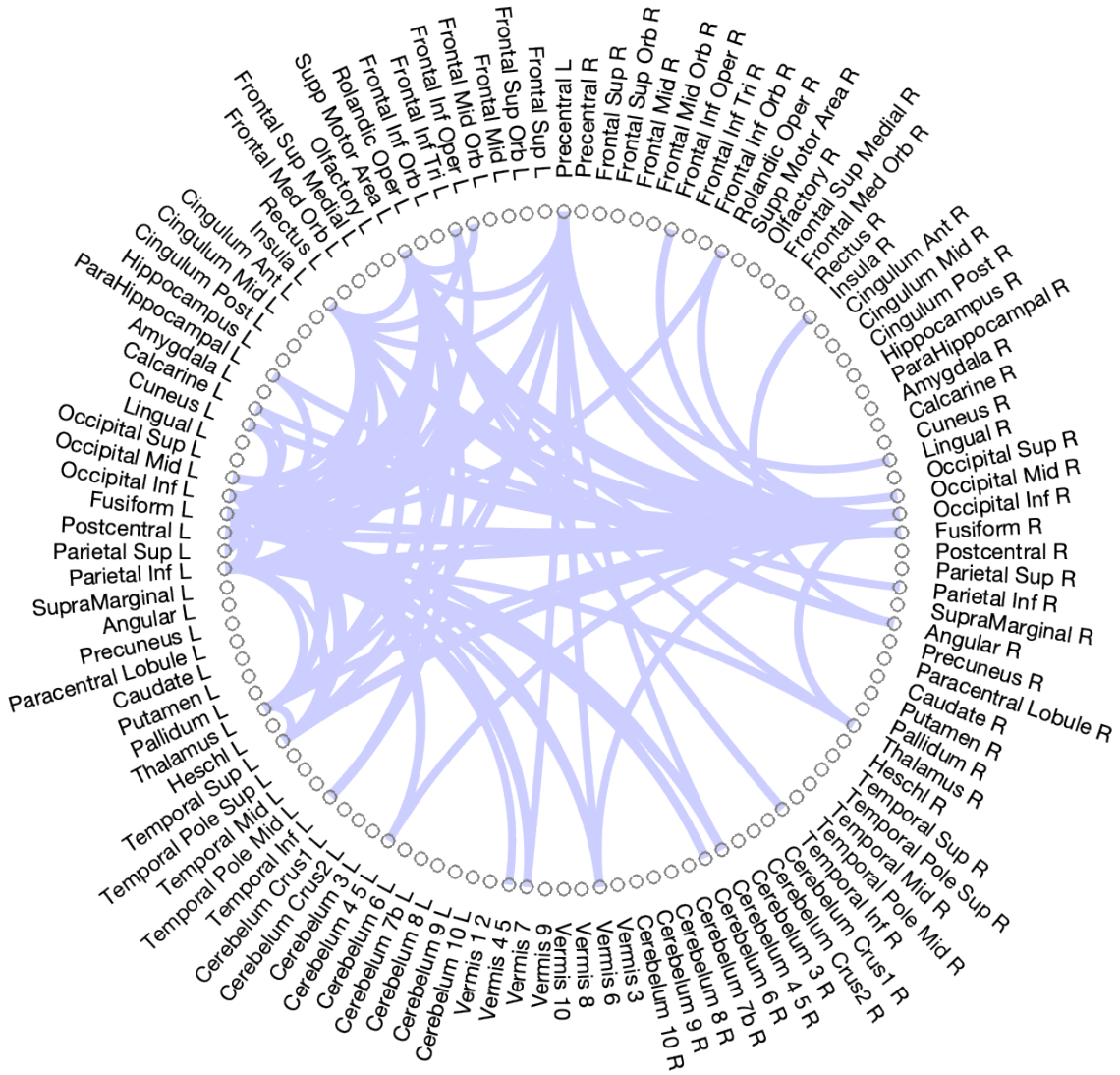


Figure 7.8: The 100 strongest connections of the mean class activation maps in resting-state/class classification for the vertical-filter model, with the maximum value taken across wavelet correlations.

While these results are still scientifically valid when analyzing whole nodes (Leming and Suckling, 2020a), it clearly affects the resolution of individual edges, which becomes problematic for fine-grained analyses.

Grad-CAM results for stochastic encoding

In total, 14,683 unique connectomes (comprising both resting-state and task data) were classified by sex across 300 ensemble models. For each connectome, a single, 116×116 gradient class activation map (with 115×58 unique values) was derived that indicated the general importance each particular edge played into the classification of that participant.

The distribution of edge values from CAMs, both from edges within, and edges connected to the respective networks, are shown for task and resting-state data in Figure 7.2. These distributions were compared to the relative distribution of all edges with aggregated values of 115×58 CAM values inside and outside of a priori networks, across 14,683 unique subjects, totalling just under 100 million values. Effect size were reported (as Cohen's d; see Figure 7.2).

The differences in CAM values of edges inside and outside the CEN were non-significant, while some effects were observed for the inner, but not connecting edges of the DMN. The largest effect was seen in the salience network, having an effect size of $d > 0.57$ for task- and resting-state data separately. In CAMs overall, there were no significant differences between task- and resting-state edge values. This likely indicates that CAMs, while useful for showing which networks are important to the overall task of sex classification, are not useful for showing whether these networks were more or less important for resting-state or task data.

Occlusion

Using the same dataset for the sex classification task, I compared the AUROCs of 300 independent models that classified a random half of the network's edges. One set of 300 deliberately included the set of edges that constituted a network, and the other set of 300 excluded the same edges (Figure 7.9, top). By comparing the AUROCs and finding a statistically significant difference, I could assess the influence of a particular network on the classification.

The relative classification (measured as AUROC) from the groupings of edges that included edges both inside and connecting to the DMN, CEN, and salience networks, as well as models completely excluding them, are shown in Table 7.1, while Figure 7.9 shows the distribution of AUROCs on 300 models including and excluding each network, for resting-state and task data.

When considering only the edges within a network (consisting of $\frac{45}{58*115} = 0.67\%$ of total edges), modest losses in accuracy were observed (Figure 7.9), but the only one that achieved statistical significance in a Mann-Whitney U-test after Bonferroni correction was the salience network classification in resting-state data. However, when excluding all edges connected to a network (consisting of $\frac{1105}{58*115} = 16.57\%$ of total edges), a difference between resting-state and task data was observed: exclusion of all three networks led to statistically significant ($p < 0.05$) decrease in AUROC for the classification of resting-state data, while the exclusion of the central executive and default mode, but not the salience networks, led to a statistically significant drop in AUROC.

7.4 Discussion

7.4.1 Deep learning model

Because it is able to capture nonlinear patterns across complex datasets, deep learning is a powerful tool for characterizing biological data. However, because of interest in identifying patterns discovered by deep learning models, the interpretability of the model is just as important as performance, though it is far more difficult to quantify or even define (Doshi-Velez and Kim, 2017). The primary methodological contribution of this study is a model that captures the contributions of individual functional connections to fMRI deep learning classification, while the results of my data show that utilisation of this model in the context of network neuroscience can shed light on between-sex differences in task- and resting-state brain networks.

My model addresses an important problem unique to the issue of classifying graphs in CNNs, which is bias inherent in its encoding. There is no universal consensus on a method of encoding graphs for ML, though others have been proposed (Jie et al., 2013; Kawahara et al., 2017; Nikolentzos et al., 2017; Kriege et al., 2019; Tixier et al., 2017; Leming and Suckling, 2020a). Whether encoding them randomly is the optimal method for classification

accuracy is up for debate, though random encoding does avoid the problem of overfitting that is present in fully-connected neural networks, and it avoids bias in the output CAMs that results from using filters with a consistent shape. In other words, the use of linear filters results in whole rows or columns of a functional connectivity matrix being emphasized, rather than particular edges. Additionally, the training scheme helped to eliminate bias from the output CAMs. Simple averaging over a large number of models and stratified cross-validation (Kohavi, 1995) is just as important as the model architecture itself, because this allows for reduced bias from both confounding factors and natural variations in the output of nondeterministic deep learning models.

Respectively, the average AUROC for sex classification across all 300 models was 0.8010; when aggregated as an ensemble, the combined AUROC was 0.8459. This represents an improvement over my previous sex classification in Chapter 4, which achieved an AUROC of 0.8295 on BioBank data (0.7683 across all datasets used) with a vertical-filter CNN balancing by only age and site. Nonetheless, due to the different balancing schemes, these two studies likely used a moderately different subset of the overall data, and so a direct comparison between the present stochastic and the previous vertical filter models in terms of accuracy is not strictly valid. Comparisons to other state-of-the-art ML studies are also not possible, since there is high variation in classification accuracy depending on how data was collected and processed (Leming and Suckling, 2020a), and few imaging studies have attempted a sex ML task on a dataset of this size.

The training and multivariate class balancing schema, when combined, offered another uniquely important contribution. By only inputting to smaller, independent models subsets of data in which measurable confounding factors were balanced beyond any detectable statistical significance, I was able to effectively regress out any confounding factors that I was able to measure. However, by combining these subsets over a large number of independent models that were then combined in an ensemble, I was able to utilize the majority of the overall data in the end result without losing the effects of balancing. This allows us to be sure that the ML model utilized the majority of an imbalanced dataset, without achieving higher accuracy due to any confounding factors, particularly head motion and intracranial volume.

Although the balancing techniques employed prevented the model from gaining higher accuracy due to confounding factors such as age, head size, and motion, this does not necessarily mean that such differences had no influence. Class balancing does not prevent the model from internally separating data based on such factors and considering them (wholly or par-

tially) independently. To illustrate this issue, I briefly present an analogy: consider a ML task in which pictures of different species of cat must be separated from pictures of different species of dog; such a model would likely identify generalized differences between each (e.g., the ear shape), while also containing internal representations of each type of cat and dog contained in the training set, relying on features unique to each individual species (e.g., stripes on a tiger). For instance, black fur color may be considered salient, even though it doesn't necessarily help to separate cats from dogs, because it helps the dataset to subclassify both black panthers and black Labrador retrievers.

Nonetheless, it is likely that class balancing within a cross-validation scheme reduced the influence of differences in confounding factors. I emphasize the importance of each particular step in the ML classification to achieve the output CAMs. These are: (1) random encoding, rather than encoding based on rows or columns; (2) averaging the output of many ML models, as individual outputs have a stochastic element; and (3) stratified cross-validation using balanced subsets of the data across these models.

7.4.2 Neuroscientific findings of CAMs from vertical filters (from Chapter 4)

When classifying sex, the model was influenced by diffuse areas connected to the frontal lobe (Figure 7.7). This is consistent with previous findings in sex comparisons of functional imaging, which did not find differences in brain activity in specific areas, but rather differences in local functional connectivity over large areas of the cortex (Tomasi and Volkow, 2011b).

Task vs rest functional connectivity classifications, as expected, identified the major components of the well-known default mode network (Raichle et al., 2001) (Figure 7.8), a set of bilateral and symmetric regions that is suppressed during exogenous stimulation (Greicius et al., 2003), as well as visual processing areas (the occipital lobe) and the supplementary motor area. Together with the comparison of sex, the confirmation of the results with those expected from the extant literature give confidence for accurate classification by the CNN as well as the specificity of the visualization method used.

The paracentral lobule and right caudate nucleus, as well as connections to the cerebellum and vermis, were identified as salient to the comparison of the autism vs TD (Figure 7.6). This finding is largely substantiated by previous studies that have found both functional connectivity and volume differences between autistic and healthy individuals in the caudate

nucleus (Sears et al., 1999; McAlonan et al., 2002; Brambilla et al., 2003; Hollander et al., 2005; O'Dwyer et al., 2016; Rojas et al., 2006; Turner et al., 2006), though these studies disagree on the exact nature of those differences (Qiu et al., 2016). Much of the literature on functional connectivity in autism, however, concerns network-wide differences (Hull et al., 2017) rather than localized differences captured by the CAMs.

The classification of autism, on average, pointed overwhelmingly to two key areas (the right caudate nucleus and the right paracentral lobule), which is consistent with many previous studies of autism. A major caveat in interpreting these results, however, is (1) their use of vertical filters, which led to bias, and (2) the use of datasets that may have been influenced by head size and motion. However, this issue is not present in the UK BioBank analysis.

7.4.3 Neuroscientific interpretations of CAMs from sex classification in UK BioBank

Four main neuroscientific findings stand out in my results: (1) when classifying sex, the relative AUROC for resting-state data was consistently higher than that for task data by a margin of around 0.12 (Table 7.1); (2) the within-network edges of the salience network were considered important for characterizing resting-state data (as indicated by both occlusion and CAM results), but not task data (as indicated by occlusion results); (3) edges connecting to all three networks were important in characterizing resting-state fMRI, and notably, even when only considering edges within the networks the p-values for differences between occlusion runs were hardly above 0.05 (Figure 7.9); (4) edges connected to the CEN were the only ones that proved important to the classification of both task- and resting-state data together (Figure 7.9), even though there was little difference in the distribution of CAM values between them (Figure 7.2).

The significantly lower classification accuracy of task data overall compared to resting-state data was consistent both when using complete input data and using partial input data (Table 7.1). The most straightforward interpretation of this result is that, in task processing, female and male brain function is more similar than it is in the resting-state. Because resting-state brain connectivity varies more than task connectivity (Elton and Gao, 2015), this disparity may also be due to a lower number of distinguishing features.

Explaining the apparent contradiction between my two methods regarding the status of the CEN is complex. Judging from the occlusion results, the CEN is an important network when

classifying resting-state data and the only network important in classifying task data, though this is not reflected in the CAMs. Given that these two methods are established visualization methods in ML and a methodical error is unlikely, the takeaway of this contradiction is that these methods are not interchangeable and must be interpreted in their own right. The contradiction could possibly be due to a relatively small number of very salient edges connecting to the CEN, which can be seen in the right tail of the histogram in Figure 7.2, though this is a very minor effect. This also shows that the interpretation of specifics in these results ought to be approached cautiously, given how novel these methods are in their application to neuroscience. Put informally, CAMs show which components of input data the deep learning model pays attention to, while occlusion shows how important a component is to the classification of a specific datapoint. With this in mind, the similar distribution of CAM values over spatially invariant task- and resting-state input data (see the histograms in Figure 7.2) is not surprising since a ML model may find a particular edge salient because it might help it to internally subclassify the datapoint by resting-state or task. Thus, CAMs may illustrate that a particular edge is important in the overall classification of the model, though not whether it helps in classifying a specific datapoint.

With regards to the salience network, however, the two methods paint a more straightforward picture, since the inner edges of the salience network were clearly the most significant, according to CAMs (Figure 7.2). Furthermore, it was the only network with inner edges that proved to be statistically significant to the classification of resting-state data (Figure 7.9). This effect may be due, in part, to the particularly salient connection between the left and right amygdala (Figure 7.2) which yielded the highest CAM value by far. The difference between men and women in amygdala response has been controversial (Andreano et al., 2014), with studies disagreeing over whether there is greater activity in men (Schienle et al., 2005; Goldstein et al., 2010; Sergerie et al., 2008) or women (Klein et al., 2003; McClure et al., 2004; Hofer et al., 2006; Domes et al., 2010) in response to affective scenes. While CAMs cannot comment on this issue, other studies have found no difference in function at all (Wrase et al., 2003; Caseras et al., 2007; Aleman and Swart, 2008), which my results refute. While I can conclude from these results that the salience network is engaged differently between males and females in, at least, the resting-state, a disproportionately high value of one of its edges may drive this classification, and thus the robustness of this results requires independent verification.

The DMN is also engaged in sex differences. As can be seen from the middle histogram in Figure 7.2, many of its inner edges have a higher class activation than other edges, while

excluding it and all edges connected with it had a uniquely negative effect on classification (Figure 7.9). What is surprising, however, is that the DMN, which is commonly cited as the marker of resting-state functional connectivity (Raichle et al., 2001) and has previously been implicated in big data sex difference studies (Ritchie et al., 2018) as an area of particular interest, does not stand out from the other two networks studied. While it is not surprising that, in the occlusion tests, the CEN had a greater effect than the DMN in task classification, both tests show that, as stated above, the salience network appears to be more important and have a greater effect on classification accuracy of the resting state. This may be due to the use of a priori tests in other studies that specifically account for the DMN, the non-inclusion of subcortical areas in other studies, or the inclusion of the critical amygdala connections in the salience network, or other unknown reasons.

7.5 Conclusion

The results of this chapter show that the distinction of males and females in resting-state takes into account all of the major brain networks, particularly the salience network, which may be as a result of increased variance in resting-state networks than task-based networks, potentially offering the model a larger set of distinguishing markers. When only considering task or, more specifically, the emotional faces recognition task of the UK Biobank, areas connecting to the DMN and, more so, the CEN showed significantly altered function, while function of the the salience network was not different enough to significantly aid in single-subject classification (Figure 7.9). Methodologically, I have also shown the applicability and limitations of two different ML visualization methods to brain network data, as well as ML's applicability to big data in a scientific field.

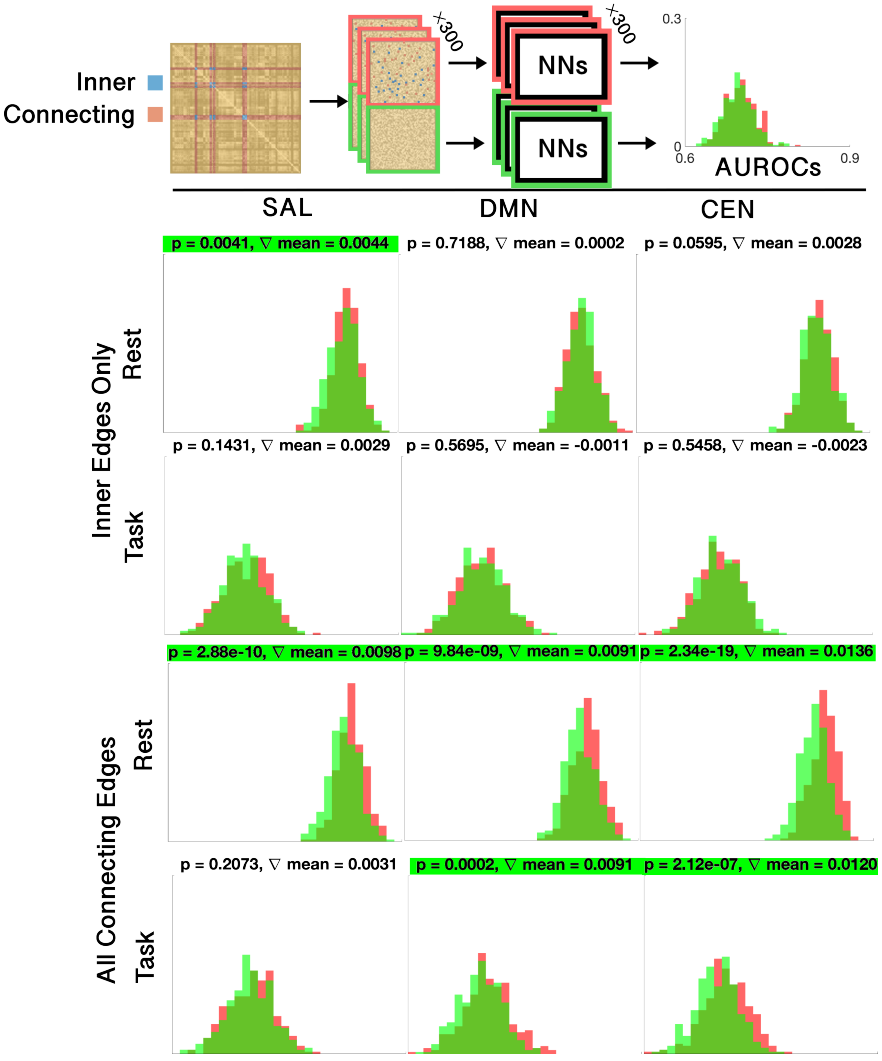


Figure 7.9: The effects of selective network occlusion on model accuracy. (Top) the process by which occlusion AUROCs are estimated; either all inner edges of a given network, or all edges connecting to a network, are selected. The network edges are then scrambled (see Figure 7.1), and the selected edges are placed among one half of the scrambled edges, and in the other half left out. These two sets are then trained on 2×300 independent neural networks, and the resulting AUROCs are compared. (Bottom) The results. Considering only inner edges, the only statistically significant effect, after Bonferroni-Holmes correction, was the salience networks on resting-state data. Considering all connecting edges, all three networks had a significant effect on the classification of sex in resting-state data, while both the default mode network and, more strongly, the central executive network, appeared to have an effect in classification of task data. The nonparametric Mann-Whitney U-test was used to test for statistical significance. Final model means and ensemble results are shown in Table 7.1.

Chapter 8

Structure/function encoding in autism

This final analysis combines the class-balancing, stochastic encoding, and visualization frameworks from Chapters 6 and 7, with the structural connectivity metric introduced in Chapter 3, for the analysis of developmental autism. I compare this method to similar classifications of the same participants using fMRI connectivity matrices as well as univariate estimates of grey-matter volumes. Further building on general themes from Chapter 3, I further applied graph-theoretical metrics on output class activation maps to identify areas that the CNN preferentially used to make the classification, focusing particularly on hubs. The results gave AUROCs of 0.7298 (69.71% accuracy) when classifying by only structural connectivity, 0.6964 (67.72% accuracy) when classifying by only functional connectivity, and 0.7037 (66.43% accuracy) when classifying by univariate grey matter volumes. Combining structural and functional connectivities gave an AUROC of 0.7354 (69.40% accuracy). Graph analysis of class activation maps revealed no distinguishable network patterns for functional inputs, but did reveal localized differences between groups in bilateral Heschl's gyrus and upper vermis for structural connectivity.

8.1 Introduction

Voxel-based morphometry (VBM) (Whitwell, 2009) is a means of detecting structural differences in brain anatomy from T1-weighted MRI across groups. In VBM, images are registered to the same coordinate space and segmented into grey matter, white matter, and CSF volumes, before comparisons are made across voxels or groups of voxels using standard statistical

tests. Due to its robustness and effectiveness, VBM has enjoyed significant popularity since it was first introduced (Wright et al., 1995; Ashburner and Friston, 2000). Structural covariance networks (Mechelli et al., 2005) correlate tissue volumes estimated by VBM in regions across groups of participants to describe relationships that are interpreted as measures of structural integrity or developmental coherence of the brain.

While there have been several cross-sectional findings of structural brain differences in autism (Redcay and Courchesne, 2005; Stanfield et al., 2008; Nickl-Jockschat et al., 2012a), these have not been substantiated by a larger-scale analysis (Haar et al., 2016). Indeed, characterizations of brain structure in autism have been inconsistent across studies of small sample sizes, although differences at different ages may explain some of this variation (Chen et al., 2011); for instance, increased amygdala volumes have been reported in children with autism (Sparks et al., 2002; Schumann et al., 2004), but not adults (Stanfield et al., 2008). A meta-analysis of VBM studies in autism found disturbance of brain structure in the lateral occipital lobe, the pericentral region, the right medial temporal lobe, the basal ganglia, and proximate to the right parietal operculum (Nickl-Jockschat et al., 2012b). Small-scale studies in children with autism have found altered structural covariance in areas associated with sensory, language, and social development. Altered structural covariance has been found between sensory networks, the cerebellum, and the amygdala in autism (Cardon et al., 2017). In children, (McAlonan et al., 2005) found that structural covariance indicated localized reductions within fronto-striatal and parietal networks and decreases in ventral and superior temporal grey matter, suggesting abnormalities in the anatomy and connectivity of limbic–striatal (i.e., social) brain system. Language ability correlated with cortical structure and covariance (Sharda et al., 2017), and associations with language development are further supported by studies showing abnormal development of the Heschl's gyrus (Prigge et al., 2013), an area where functional activation has been associated with development of 'inner speech" (Hurlburt et al., 2016). In adults with autism, structural covariance has shown decreased centrality in cortical volume networks (Balardin et al., 2015).

Functional connectivity in autism has previously been discussed in Sections 1.6.3 and 4.1 of this thesis. Briefly summarizing, autism has been consistently associated with differences in brain function (Simas et al., 2015a; Müller et al., 2008). Efforts to find differences in functional connectivity relative to neurotypical control groups have characterized autism as exhibiting under-connectivity, and thus greater segregation of functional areas (Just et al., 2004; Cherkassky et al., 2006; Kennedy and Courchesne, 2008; Assaf et al., 2010; Jones et al., 2010; Weng et al., 2010). Other studies, mostly of children and adolescents, found

evidence of over-connectivity in specific areas of the brains of those with autism (Cerliani et al., 2015; Chien et al., 2015; Delmonte et al., 2013; Di Martino et al., 2011; Nebel et al., 2014a,b), locating hyperconnectivity to the posterior right temporo-parietal junction (Chien et al., 2015) and in striatal areas and the pons (Delmonte et al., 2013; Di Martino et al., 2011). Hull et al. (2017) posited that autism is likely characterized by a mix of hyper- and hypo-connectivity traits.

Section 1.6.3 also covers previous efforts in machine learning at whole-brain classification in autism, though the most relevant to the present study is Eill et al. (2019), which compared autism classification from brain structure and brain function by performing a classification on individuals with autism and neurotypical controls using structural MRI, DWI, and fMRI data, finding that features derived from fMRI provided the highest accuracies with an SVM classifier. They did, however, encounter the issue of fMRI feature extraction simply producing more variables than its structural counterparts, offering the machine learning model more information to work with, although attempts were made to mitigate this issue.

8.1.1 Studies of the structure-function relationship in the brain

The relationship between brain structure and function has long been of general interest and study in neuroscience, encompassing several decades of research in itself . In the context of MRI specifically, it is generally approached in one of three different ways: computational modeling, correlating structural and functional measurements, and studies of brain lesions.

The first comparison of structure and function in MRI was Koch et al. (2002), which found little evidence of correlation within the one axial slice of the brain examined. Later, following the emergence of interest in resting-state fMRI, Greicius et al. (2009) utilized DTI tractography to study resting-state functional connectivity, that structural connectivity did not always predict resting-state. Hagmann et al. (2008); Honey et al. (2009) identified a structural core in the parietal lobe around the area of the default mode network that was found to strongly predict the presence of functional connectivity; Horn et al. (2013) found that voxel-by-voxel structural-functional coupling was particularly high in the default mode network. While systems with high structural connectivity produce more reproducible patterns in functional connectivity (Honey et al., 2009; Damoiseaux et al., 2012), Greicius et al. (2009) established that functional connectivity does not necessarily imply structural connectivity, and there have been a number of highly connected areas of the brain with no apparent linking white matter tracts, which has been a driving question behind the structural-function

relationship (Vincent et al., 2007; Skudlarski et al., 2008; Honey et al., 2009; Adachi et al., 2012).

Comparison of the structural and functional connectomes has found many of the same topological organizations, such as rich-club, modularity, and small-world properties (Wang et al., 2015) (though it is unknown whether these properties are unique to brain networks or present generally in real-world graphs). Direct correlation of structural connectivity strength and resting-state functional connectivity have reported R-values between 0.48 and 0.78 (Hagmann et al., 2008; Honey et al., 2009; Wang et al., 2015), with variation depending on subjects and the selection and method of parcellation. Correlations between structural and functional connectivity are further complicated by the natural variation in brain function. While it is known that people have distinct, individual functional patterns (Finn et al., 2015), functional connectivity changes based on attentional demands (Hermundstad et al., 2013), in response to learning (Bassett et al., 2011), and dynamically just in response to time spent in an MRI scanner (Chang and Glover, 2010; Hutchison et al., 2013; Allen et al., 2014).

The study of structural-functional relationships is also rooted in theoretical neuroscience, particularly simulation or prediction of brain function based on a structural underpinning. Models of both human brain networks (Kaiser and Hilgetag, 2004; Kaiser et al., 2007) and functional human brain dynamics (Alstott et al., 2009) that have simulated lesions in the structural connectome, have found that damage to high-centrality structural nodes disrupts functional connectivity the most, both in the direct vicinity of the lesion and in remote brain regions. Nonlinear simulations of fMRI data, such as the use of Kuramoto oscillators (Schmidt et al., 2015) with an input structural substrate, shed light on the role of local and global neural dynamics in spontaneous brain function (Ghosh et al., 2008; Deco et al., 2009, 2011, 2013; H utt et al., 2014).

Real-world results on human patients with lesions have supported these results; Johnston et al. (2008) found inter-hemispheric functional connectivity to have disappeared in a young patient following a callosotomy, while intra-hemispheric functional connectivity was unaltered, though this result is more ambiguous in other studies that have found inter-hemispheric connectivity can widely be preserved following callosal agenesis or surgical lesions of the corpus callosum (Uddin et al., 2008; Pizoli et al., 2011; Tyszka et al., 2011). He et al. (2007) studied patients with a stroke in their right hemisphere, correlating impairment of attention with disruptions in the attentional network outside of the stroke area.

Questions of the relationship between structure and function has been applied to specific

phenotypic differences as well; a review by Batista-García-Ramó and Ivette Fernández-Verdecia (2018) noted that the four major areas of interest in this regard are aging, epilepsy, schizophrenia, and autism. Early childhood studies of autism have reported increased structural connectivity in young children and adolescents (Ben Bashat et al., 2007; Chang and Glover, 2010), which may be related to decreased inter-hemispheric anatomic (Lo et al., 2011; Weinstein et al., 2011) and functional (Anderson et al., 2011a) connectivity; however, this may also be a symptom of other reported differences in functional networks in autism, such as average path length and modularity (Rudie et al., 2013; Simas et al., 2015a).

8.1.2 Experiments

In the present work, I applied the structural connectivity matrix estimation method from Chapter 3 to the deep learning methods developed throughout this thesis. While the structural connectivity metric has no hypothesized physiological interpretation, it acted as an effective means of dimensionality reduction that allowed for T1-weighted MRIs to be encoded into a machine learning model. I then compared this to classification using only functional connectivities, as well as univariate gray matter volume estimations. I then derived class activation maps (CAMs) from all of these data. I used the output class activation maps (CAMs), combined with graph theoretical techniques, to understand which parts of the brain the model focused on, and whether simple linear regression models could spot the same qualities in these data. To study the structure-function relationship, I further describe a means of combining the structural connectivity matrices with functional connectivity matrices in the same machine learning model to yield improved accuracy. Last, I show how the classification results differ across different age groups.

8.2 Methods

8.2.1 Dataset

As stated in Chapter 3, I used a dataset comprised of 29,288 total instances each with a structural MRI and a functional MRI in both task-activated and task-absent (rest) conditions. (Note that in many instances, data were acquired from the same participant.) In total, 1555 data points were from participants with autism. These data were drawn from

					Age				Sex		
Collection	# Subj.	# Conn.	Rest	Task	Min	Max	Mean	Stdev	Female	Male	Autism
ABCD	1049	5142	2296	2846	0.42	11.08	10.12	0.69	2474	2668	61
ABIDE	412	412	412	0	6.00	45.00	17.00	7.16	45	367	181
ABIDE II	682	717	717	0	5.22	55.00	14.39	7.39	169	548	350
BioBank	9791	9791	9791	0	40.00	70.00	55.00	7.51	5178	4613	4
NDAR	1050	7958	5531	2427	0.58	55.83	18.71	7.80	3816	4142	930
Open fMRI	1194	5268	820	4448	5.89	78.00	27.12	10.24	2346	2479	29
Total	14178	29288	19567	9721	0.42	78.00	30.72	–	14028	14817	1555

Table 8.1: Statistics for each dataset used.

six different databases: OPEN fMRI, the UK BioBank, ABIDE I, ABIDE II, NDAR (minus ABCD), and ABCD (Table 8.1). Covariates of age, sex, task were also compiled.

8.2.2 Pre-processing and feature extraction

The full pre-processing pipeline for functional data is described in Chapter 2. To re-iterate: functional data were preprocessed using SpeedyPP. Data were first skull stripped. Motion was regressed from time series using wavelet despiking (Patel et al., 2014). Data were then registered to the stereotatic space of the Montreal Neurological Institute (MNI), after which they were overlaid on the 116-area AAL parcellation. Datasets with greater than 10% regional dropout, or which otherwise failed the SpeedyPP stage, were excluded. The remaining datasets are presented in Table 8.2. 116×116 functional connectivity matrices were estimated using Pearson correlation on the averaged timeseries within a region.

To estimate grey matter volumes of each area in the AAL parcellation, structural MRI were first skull stripped using tools from the Analysis of Functional Neuroimages (AFNI) toolbox, then registered to MNI space and grey matter values estimated using FSL VBM. Grey matter volume estimations at each voxel were then averaged within the areas of the AAL parcellation, producing a 116×1 array.

8.2.3 Machine learning model and training

I classified individuals with autism and neurotypical controls using, separately, structural connectivity, grey-matter volume, and functional connectivity measurements, as well as a

model that combined structural and functional connectivities. I employed the model and training scheme described in Chapters 6 and 7. This used an ensemble of 300 convolutional neural networks that each scrambled the unique values of input connectivity matrices, losing some spatial encoding information while avoiding biases in output class activation maps (described below).

In building training, test, and validation sets for the models, the multivariate class balancing scheme was used. Equal ratios of autism and neurotypical control participants were enforced, and equal distributions of age, collection, and intracranial volume were maintained in each class. To account for motion, equal distributions of mean framewise displacement of fMRI data were also maintained, but, unlike in Chapter 7, DVARS and mean spike percentage were excluded from the class balancing scheme, as a way to increase training set sizes. The class balancing scheme divided data into test, training, and validation sets for each model in the ensemble, ensuring participants with multiple functional connectomes were in the same group. Each model was trained on an Adam optimizer for 100 epochs, after which the epoch with the highest accuracy on the validation set was used. This model then made a prediction on each instance in its test set.

An ensemble of 300 independent CNN models was used to make predictions on the same test set, and an AUROC derived by averaging across instances. When adding models to the ensemble, the AUROC from the aggregated models increased in a predictable way. The AUROCs from between 20 and 300 models were fit to a logarithmic curve with a hard limit in order to predict the projected highest AUROC possible in the limit of a large number of models.

As a result of forced class balancing, each model in the ensemble used an independent subset of approximately 1600 instances. As an effect of this balancing scheme, data from Open fMRI and the UK BioBank, having few participants overall with a diagnosis of autism, were included only infrequently, while data from ABIDE I and II, ABCD, and NDAR were frequently represented.

In total, four cross-sectional classification tasks were undertaken (Table 8.2), specifically: with structural connectivity; with grey matter volumes; with functional encoding; and by combining structural and functional connectivities.

Modality	AUROC	Accuracy
Structural conn., Function	0.7354	69.3980
Structural conn.	0.7298	69.7062
Function	0.6964	67.7180
Structure (GM vols)	0.7037	66.4228

Table 8.2: Respective AUROCs and accuracies of ensemble models on different combinations of data.

8.2.4 Class activation map analysis

Using the Guided Gradient Class Activation Map (Grad-CAM) algorithm (Selvaraju et al., 2017), which displays areas of the input data most salient in classification, I measured the class activation of each data point in each model proposed, and then averaged these maps generating a 116×116 CAM for both structural and functional connectivity, as well as a 116×1 map for grey matter volume. I correlated the structural and functional CAMs to the measured effect size of differences between autism and neurotypical controls for the connectivity data, as a way to determine the similarity of CAMs to conventional statistics.

Next, I isolated hubs in the 116×116 CAMs. To do so, I first measured the edge betweenness centrality of each edge in the CAMs. I then grouped these values into different communities by maximizing modularity of the edge betweenness values (Brain Networks Toolbox (Rubinov and Sporns, 2010)). This procedure identified which hubs were most focused on by the classifier.

8.3 Results

8.3.1 Training

Training accuracies and AUROCs are given in Table 8.2.

Classification resulted in a higher AUROC for structural than functional connectivities: 0.7298 and 0.6964, respectively. Classification on univariate grey matter volumes resulted in an AUROC of 0.7037, outperforming functional classification while underperforming structural connectivity classification, although this might be expected considering its lower di-

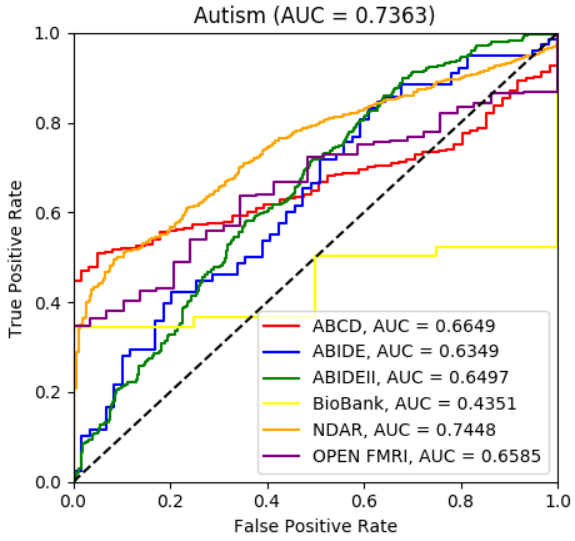


Figure 8.1: Overall AUROCs of each dataset included in the analysis for the structure/function ensemble.

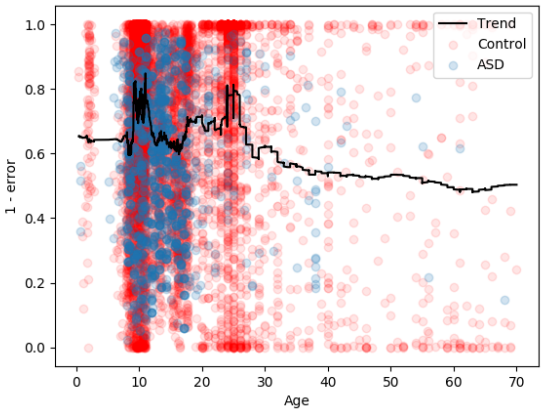


Figure 8.2: Relative classification error in the structure/function/age ensemble model, plotted against age. Each point in the graph represents the averaged classification error of the datapoint across each model in which it was included in the test set. Thus, more controls are represented which were each used individually in fewer models, while the autism datasets are fewer but were generally used in more models. Thus represents that accuracy was generally higher in the developmental age groups, likely because more data was present for those groups.

mensionality. Combining structure and function resulted in an AUROC of 0.7354 (Figure 8.1, left), with a projected upper limit of the AUROC of 0.745 (Figure 8.1, right).

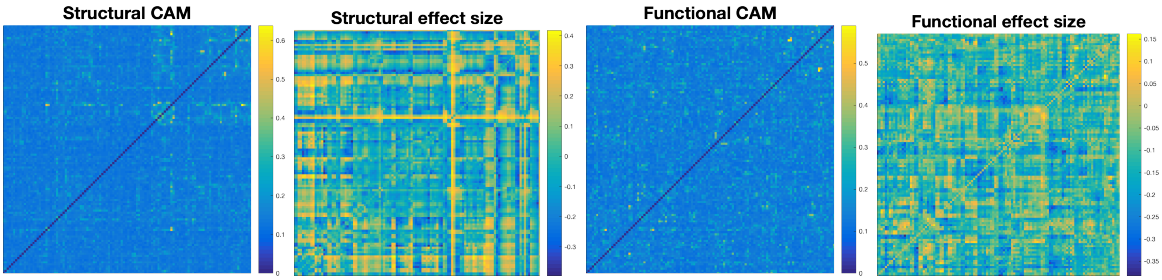


Figure 8.3: A comparison of the effect size of differences between raw matrix values between groups and the averaged class activation maps. Most of the edge differences passed a non-parametric statistical significance test. When comparing the CAM matrix and the effect size matrices using either linear or nonparametric correlation, neither had any statistically significant associations with one another.

Figure 8.2 shows the classification results across different age groups, reflecting the large disparities in age ranges present in the accumulated dataset, as well as the heightened model performance for those age ranges for which the most data was present. This reflects both the disparities in autism characterization across development as well as the likelihood of increased accuracy with more heterogeneous datasets.

8.3.2 Class activation map analysis

When comparing the output CAMs to their respective functional and structural effect sizes, no statistically significant correlation was observed, and thus the machine learning model relied very little, if at all, on differences detectable by conventional statistics (Figure 8.3).

CAMs for structural and functional connectivities, sorted by different detected communities after edge betweenness centrality was measured, are shown in Figures 8.4 and 8.5. Structural CAMs showed five distinct groupings, each with distinct hubs that each centered on one or two localized areas, including the left and right Heschl's gyrus, the upper vermis, the right frontal-medial orbital gyrus, the right pallidum, and the left putamen. The strongest activations were found in left Heshcl's gyrus.

Localisation was also found, though less distinctly, in functional hubs, notably the left inferior parietal lobe, the left middle temporal lobe, the left olfactory bulb, and the upper vermis. However, focus on particular hubs was not a distinctive feature.

CAMs for grey matter volumes are shown in Figure 8.6. These results had very little in

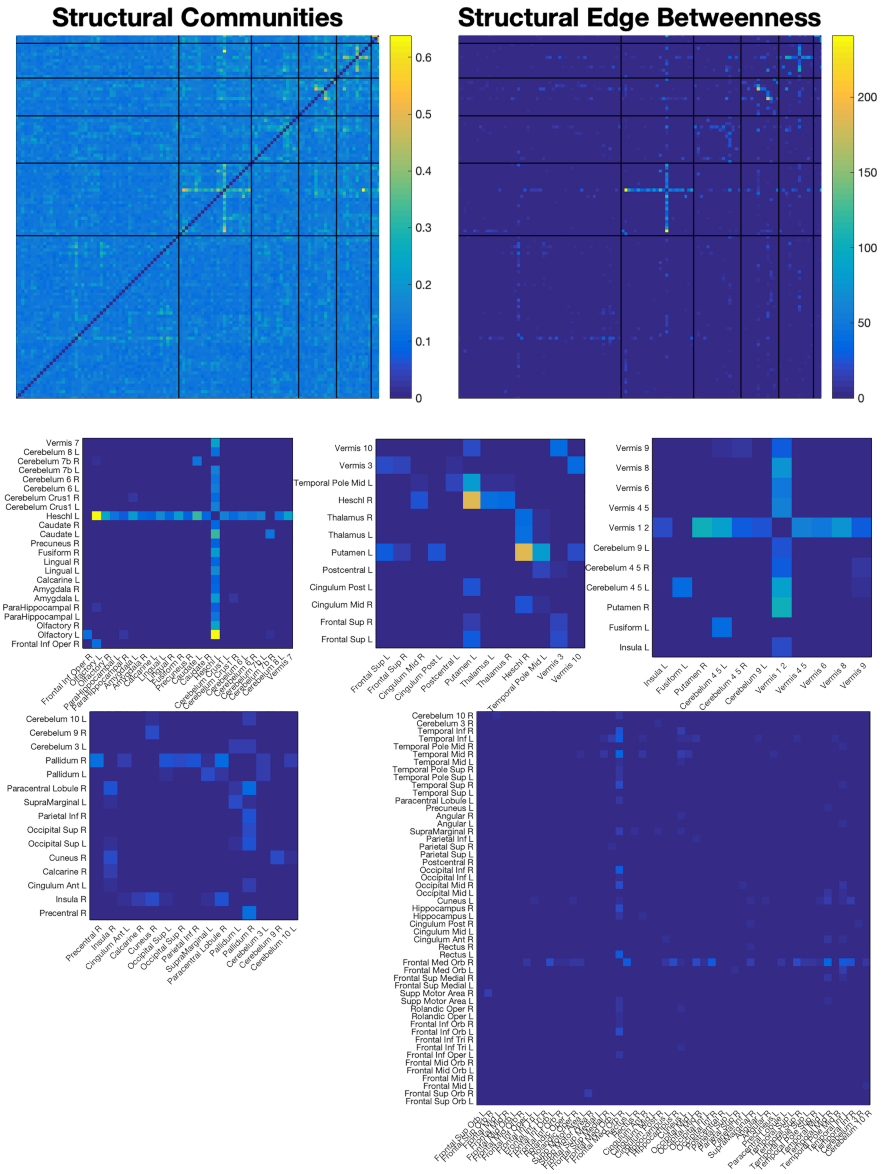


Figure 8.4: The structural hubs targeted by the structure/function/age encoding. Shown here are the class activation maps (upper left) as well as the edge betweenness centralities of the map (upper right), after it as been sorted into six different hubs via modularity maximization. The hubs, with labeled areas, are shown in the bottom half. (Middle) The three most distinct hubs revolve around the left Heschl's gyrus; the right Heschl's gyrus (and, to an extent, the left Putamen); and the upper vermis. The largest hub, in the bottom left, shows scattered-but-weak emphasis on connections to the right frontal medial orbital gyrus. These connections likely reflect the machine learning model's use of comparisons of certain areas to others in order to assess the developmental difference of such areas in autism.

common with the structural connectivity results, with the strongest five activated areas in the right supplementary motor area, the right middle frontal lobe, the right precentral sulcus, the left insula, and the inferior frontal gyrus triangularis.

8.4 Discussion

In the univariate grey matter volume results, the CAMs highlighted the right supplementary motor area, right mid frontal lobe, right precentral sulcus, left insula, right frontal inferior triangularis, left frontal inferior orbital lobe, and the right superior temporal lobe (the top 20 areas are shown in Figure 8.6). Comparing the CAM emphasis of the grey matter volumes to the meta-analysis of autism VBM studies in Nickl-Jockschat et al. (2012b), which found six areas with consistently altered grey matter volumes, some similarities can be seen, notably in the right superior temporal lobe where grey- and white-matter volume differences in the right medial temporal lobe and the left post central gyrus.

Functional analysis did not reveal a pattern of local hubness characterizing structural connectivity differences, but rather focused on specific connections. However, a number of general functional communities were identified (Figure 8.5). Meta-analyses of studies in functional connectivity differences associated wtih autism have not found consistent differences in the brain, but rather in network-wide measures (Hull et al., 2017). The lack of hub emphasis in functional results may be additional evidence of network-wide, rather than localized differences between autism and neurotypical control groups seen in other recent findings (Suckling et al., 2015).

In structural connectivity, three definitive hubs were identified: left Heschl's gyrus, right Heschl's gyrus, and the upper vermis. The right pallidum and fronto-medial orbital region also showed relatively strong local hubs, though to a lesser degree. Emphasis of the Heschl's gyrus is in agreement with recent studies in developmental autism, having been implicated previously as an area that develops atypically in autistic children (Prigge et al., 2013). Function of the area has been associated with development of "inner speech" (Hurlburt et al., 2016), indicating a difference in development of language capabilities. My findings differ in that they found this emphasis in *structure* and not *function*, but this may be reflective of the lower variability of differences across a single area in the development of grey matter as opposed to function, which likely varies far more across participants, and age groups. The cerebellum, meanwhile, has consistently been cited as an area of difference between

individuals with autism and neurotypical controls during development (Chen et al., 2011), as well as an area of difference in structural covariance associated with autism (Cardon et al., 2017).

The structural connectivity CAMs resulting from this study revealed an emphasis on a number of distinct and localized areas, and these areas were clarified by use of an edge centrality measurement combined with modularity maximization to isolate hubs. The edge betweenness step was added by necessity to place extreme emphasis on a smaller number of more central edges, and only then could modularity maximisation isolate hubs in a meaningful way (see Figure 8.4).

The structural connectivity method's efficacy with the machine learning model suggests that it encoded practically useful information about brain structure, but the interpretation of what these structural hubs indicate physiologically is more complicated. While some correlation is present (Figure 3.14) the functional and structural connectivities show largely different patterns. Furthermore, considering that the method used to estimate structural connectivity was a similarity metric, the emphasis on these hubs was less likely an indication that they were centers of a physiological brain network characterizing autism. Because the strength of connections was a comparison of grey matter distributions, it is more likely that connections to the identified hubs were used by the machine learning algorithm as a proxy for detecting subtle changes in the morphology of grey matter within those specific regions. Edges connecting to these structural hubs were probably an indirect indication of differences in grey matter between two areas, and the individual connections themselves would not indicate any special physiological relationship. However, this still means that the hubs themselves were important in characterizing autism. This lack of an explicit physiological interpretation of this metric, however, does not detract from its utility in the context of machine learning. This structural connectivity metric may simply be viewed as a way of encoding relative spatial information about the morphology of individual areas of the brain.

The univariate grey matter volume results further complicate interpretation because areas different from the structural connectivity results were emphasized by the CAMs, even though both univariate grey matter volumes and structural connectivities were derived from the same imaging data. This brings up three key points. First, the method of encoding data is important because it presents different types of information to the machine learning model. Structural differences in autism (and likely other phenotypic differences) may vary in different ways that are only apparent under specific methods of encoding, and thus the model may have focused on different areas, depending on which method of encoding was performed. This

is important for both interpreting the results in the context of a specific machine learning task and understanding the underlying physiological implications. Second, the emphases presented by Grad-CAM were *relative*; that is, in analysing the distribution of Grad-CAM values, I saw that the model took all areas into account (Figure 8.6), although with highest focus on the few areas that seemed to hold more influence in the final classification task. This does not, however, mean that other areas were ignored entirely. Third, because of the higher dimensionality of structural connectivities over grey matter values, it may be the case that the machine learning model assumed information about grey matter volumes from a small number of edges, while information about differences in morphology of other areas (e.g., the left and right Heschl's gyrus), which were not present in the univariate feature extractions, required emphasis by a greater number of edges; this may be crucial to understanding differences in autism generally, or it may have simply helped the model increase AUROC by a margin of 0.0891 between the univariate and connectivity classification tasks. Stated informally, differences in morphology detected by the structural connectivity matrices were more subtle, and so they required the emphasis of a larger number of edges.

The results in Figure 8.2 show that autism classification, even when structure and function are considered, generalizes poorly across large age groups. This supports the findings in recent longitudinal studies of autism (Ha et al., 2015; Lange et al., 2015; Wolff et al., 2018), which found high inter-individual variability in brain volume growth trajectories in autism. This suggests that autism is highly variable in its development and that information about one age group with autism would not necessarily inform predictions of another age group.

It is notable that none of the classification accuracies presented in this paper approached the success required for a clinical diagnosis, which would need to consistently exceed 95% accuracy on a substantially large dataset. A likely reason for the comparatively low accuracy in this study specifically is the large dataset size, which, in the context of whole-brain MRI classification, has previously been associated with a drop in accuracy (Katuwal et al., 2015; Arbabshirani et al., 2017). Nonetheless, deep learning models are useful in these contexts both as statistical models in and of themselves to study autism, and as building blocks to approach clinical-quality accuracy in the future.

Finally, I combined my structural connectivity metric with functional connectivity raising the final AUROC. This shows that my method does not have to be considered as a replacement for any previous methods, but may be used in combination with them in order to make single-participant classifications more effective.

8.5 Conclusion

The present work offers a means of encoding T1-weighted MRI for use in network-based machine learning models, and with a machine learning classification task I have demonstrated an increase in accuracy in classifying individuals with autism when compared with both functional connectivities and classification of univariate grey matter volumes. Furthermore I presented methods of identified areas emphasized by the machine learning model, demonstrating the importance of data encoding and highlighting complications with interpreting results when the feature extractions have no specific physiological interpretation. While this tradeoff, interpretability for higher accuracy, will likely continue to be an issue in machine learning with scientific data, the effects of data encoding on accuracy point towards feature extraction methods as a future direction of investigation.

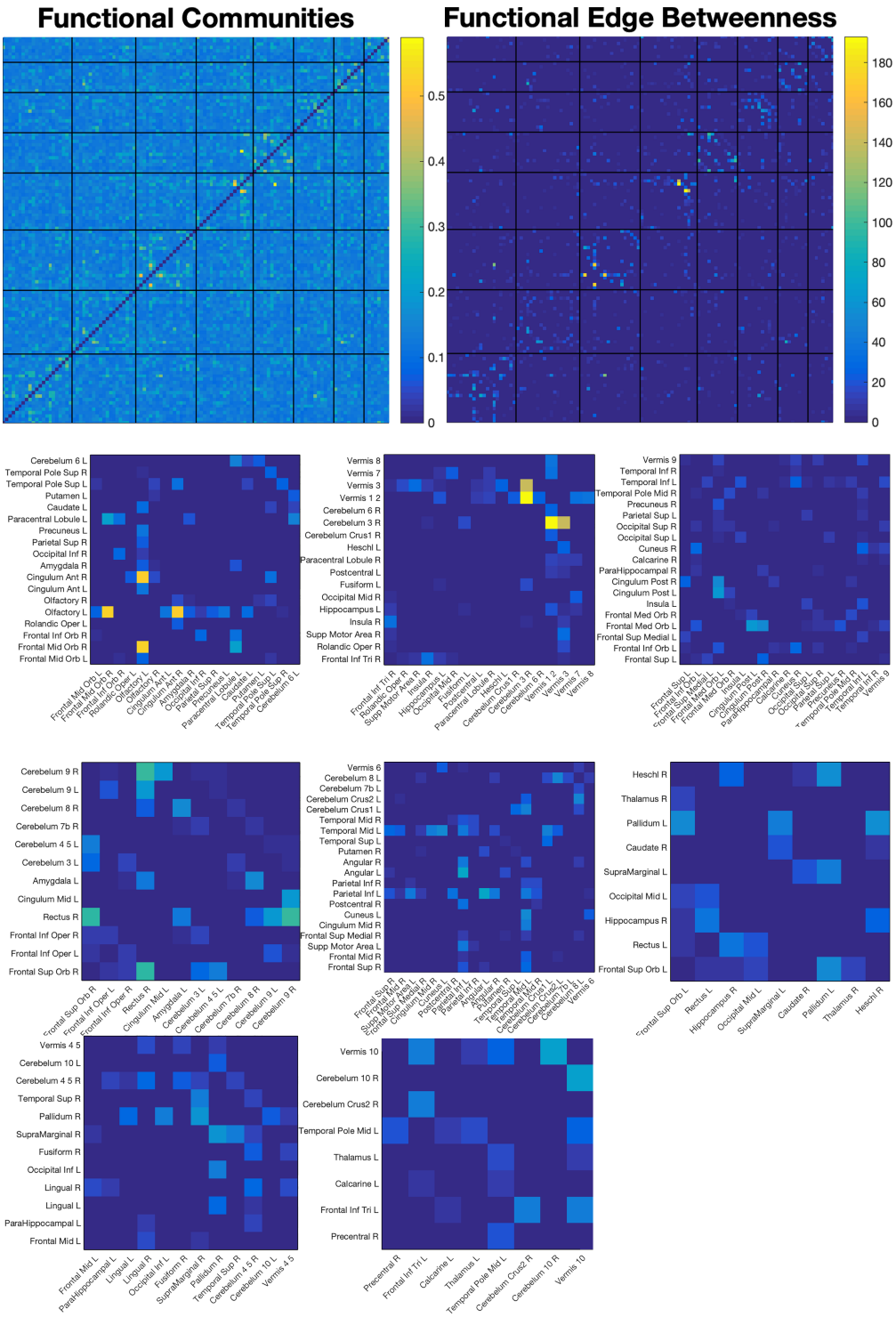


Figure 8.5: Averaged functional class activation maps and the associated edge betweenness centralities, when divided into communities via modularity maximization. Function does not show the same ultra-localized hubness within particular communities in the way that the structural results do, but emphasis is given to several individual connections throughout.

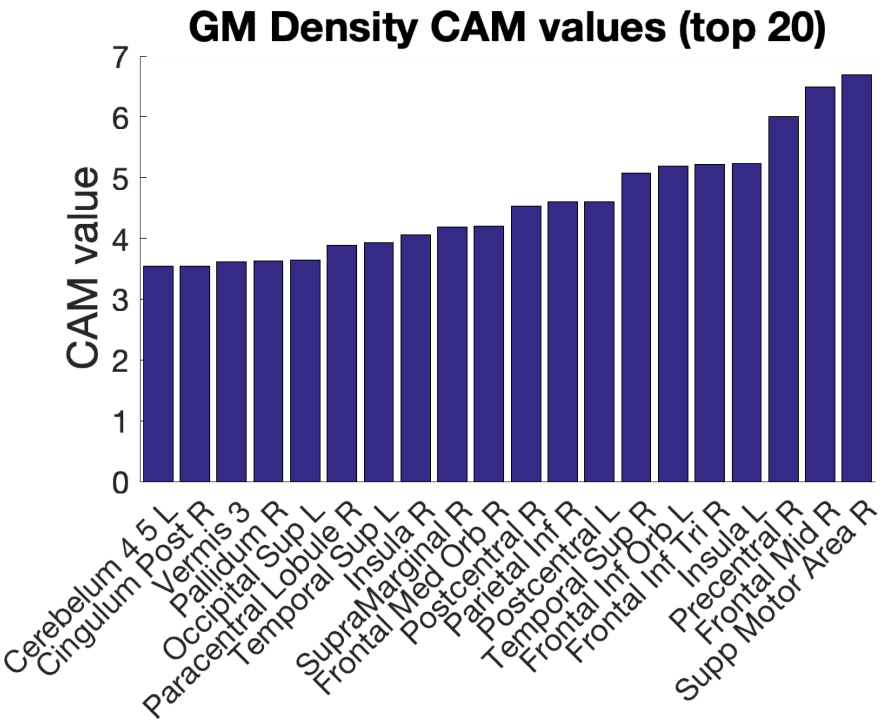


Figure 8.6: Top class activation map value results for the 116-area gray matter density classification, showing the areas most focused on in that classification task. The minimum CAM value (not shown) was 1.3622.

Chapter 9

General discussion

9.1 Summary

While several chapters of this thesis presented their own contextual neuroscientific results, the overarching contribution of this thesis is methodological. To summarize, the following were analyzed:

- Chapter 2 described the acquisition and preprocessing of a large, mixed-site dataset, as well as the practicalities and caveats of the presented deep learning schemes.

- Chapter 3 described two methods of analyzing brain connectomes: the first was a method of finding average shortest pathways in a group of functional connectomes, and the other was a method of deriving structural connectivity from T1 MRIs. This was as an initial exploration in the use of graph theoretical metrics for analyzing mental disorders.

- Chapter 4 described the first deep learning study, which used ensemble CNNs with vertical filters to classify functional connectome data by sex, rest/task, and autism. This was an initial foray in this territory; future chapters fixed issues with this study and expanded on it in different ways.

- Chapter 5 described a method of analyzing the clustering of datasets throughout the ensemble from Chapter 4.

- Chapter 6 described a multivariate class balancing algorithm to apply to the ensemble scheme and address some of the issues pointed out in Chapter 5.
- Chapter 7 described different techniques for determining the saliency of edges and networks in functional connectomes, then applied these to sex classification in rest/task data in the UK BioBank.
- Chapter 8 used the structural connectivity metric from Chapter 3, as well as functional connectomes, to classify by data by autism. This also presented a more sophisticated use of graph theoretical metrics to analyze salience in class activation maps.

Thus, throughout this thesis, I have developed a comprehensive machine learning paradigm to study large numbers of structural and functional MRI datasets.

Differences between the setup of the studies in Chapters 3, 4, 7, and 8 were highly dependent on the context of the studies themselves, as well as the new methodological innovations presented throughout. For instance, the question of adolescent depression would have been very interesting in a machine learning context, but the MR-IMPACT dataset from Chapter 3 was too small for application to a deep learning model. Likewise, the algorithm invented for estimating normative pathways from the first study relied on combinatorics, making it inapplicable to the substantially larger dataset used throughout. The choice to reduce from four to three wavelet correlation frequency bands between Chapters 4 and Chapter 8 was made to increase the number of autistic datasets included in the study, as too many were eliminated by their TR rate. The choice to only analyze UK BioBank for sex classification in Chapter 7 was made to reduce variation in the number of tasks considered in the task fMRI data. And the choice in Chapter 8 to switch to Pearson correlation from wavelet bands was done so that functional connectivity would not contribute more information as an input to an ML model than structural connectivity, and their relative contributions to accuracy could fairly be compared. Additionally, the innovations of Chapters 6 and 7 with class balancing and stochastic encoding made the CAM results more reliable than those from the study in Chapter 4, which employed vertical and filters with only age- and collection- and age-based balancing.

Nonetheless, such changes do make the results of these studies more difficult to compare, as any one of the discrepancies between the results may be explained by changes in methodology.

The study of normative pathways in Chapter 3 marked a departure from the rest of this thesis, which is focused on deep learning. However, this study is included for three reasons:

first, as an initial study in the analysis of group functional connectomics, which represents my first year of work; second, it is tangentially related to the use of graph theoretical metrics used in Chapter 8; third, it is possible that the advancement of machine learning models in the future that are capable of whole-graph classification connectomes will rely on pathfinding to some degree as an encoding or feature extraction method, and normative pathways may well play a part in this (however, the development of such a new paradigm in deep learning is beyond the scope of the present work); this idea is expanded upon in Section 9.9.

9.2 Big versus small datasets

A phenomenon that has been noted in literature on whole-brain classification is the disparity between classification accuracies between small-sample-size studies (which routinely achieve $> 90\%$ classification accuracies (Arbabshirani et al., 2017)) and larger datasets, such as the one presented here, as well as the studies of databases such as ABIDE (Khosla et al., 2018). This tendency has previously been noted in the context of structural MRI autism classification (Katuwal et al., 2015), as well as SVM studies in medical imaging (Arbabshirani et al., 2017). This phenomenon apparently contradicts the conventional wisdom of machine learning, which is that larger training datasets lead to higher classification accuracy on a test set. Given that such smaller datasets are often collected and analyzed internally by individual groups, several factors may contribute to this phenomenon: poor motion regression in the pre-processing stage, site differences or differences in MRI scanners, differences in data quality, inconsistent diagnostic metrics, access to metadata, the ability to curate the dataset during the recruitment stage, and group homogeneity within local geographic areas in which recruitment takes place. However, some of these reasons are only applicable to certain classification tasks (for instance, diagnosis of autism varies more than determination of biological sex).

One study (Schulz et al., 2019) directly tested this phenomenon on sex classification in the UK BioBank using a variety of machine learning models, finding, in that context, that accuracy did rise with the size of a training set. Thus, this issue may just be limited to large, mixed-site datasets. However, the ensemble projection results from Chapter 4 show a much smoother increase in accuracy with sex classification with the other two classification tasks, so differences in the stability of model performances may be related to the nature of the classification task itself.

9.3 Psychiatric diagnosis in machine learning

A clear limitation and area of expansion in the methodologies presented in this thesis concerns the use of binary classification. For classification tasks such as biological sex, binary classification is more-or-less appropriate, but the diagnosis of mental disorders (including and especially autism) is most often defined in multiple subcategories or on a spectrum. A more appropriate formal definition of autism for a machine learning problem would include both sub-categorizations and continuous variables indicating severity. In the context of multi-site big data, I was limited by available data; while autism is most often defined on a spectrum, available data only provided binary classifications between autism and controls. This ill-defined machine learning problem may be a reason for the underperformance and instability of autism classification compared to sex classification.

Furthermore, while brain function and structure has been associated with mental disorders, such disorders are usually defined by their outward symptoms and may have no discernable connection with brain function and structure at a macro level. For a set of mental disorders, it is thus unlikely that a machine learning model, having knowledge only of physiological measurements of the brain and not of outward symptoms, can achieve near-perfect classification accuracy (Borsboom and Cramer, 2013), even with stronger deep learning models and higher-resolution data than that presented in this thesis. While deep learning for MRI has a high potential for clinical diagnosis and prognosis, it is probable that it will only be applicable to a certain number of mental conditions that have strong biomarkers. However, further research is needed to know which mental conditions these are.

9.4 Class balancing techniques

This thesis presented a class balancing technique as a means of regressing confounding factors from data. I presented this as an optimization problem that seeks to find a subset of data, such that the distributions with respect to any one particular component were not detectable with any statistical significance, while maximizing the amount of data present in the subsample. I reiterate the formalization of this problem described in Chapter 6: if S is a nonparametric, two-sample test for statistical significance, such that $S(A, B) = 1$ if the null model can be rejected with statistical significance and $S(A, B) = 0$ if it cannot; $A = a_1, a_2...$ and $B = b_1, b_2...$ are datasets A and B, $a_i^{(j)}, 0 < j < J$ and $b_i^{(j)}, 0 < j < J$ one of

J (continuous or discrete) measurable confounding factors of the datapoints, then the class balancing problem seeks to optimize the following:

$$\underset{|A'|}{\operatorname{argmax}}(A' \subset A, B' \subset B \mid |A'| = |B'| \wedge \sum_{j \in J} S(A'^{(j)}, B'^{(j)}) = 0)$$

Note that, for discrete variables, S would simply ensure that there are equal numbers of either in A' and B'.

Though a heuristic-based approach for this optimization was described in Chapter 7, this thesis did not claim that the method presented offered the global maximum, but simply one that could be practically computed; this is even more true of the quartile balancing used in Chapter 7, which balanced between sexes and resting-state/task, and for which the method was to simply apply the algorithm from Chapter 6 twice. Improvements in methods for finding balanced subsets of two classes would be important not only because they would allow individual deep learning models to use higher sample sizes, but, because more confounding factors would inevitably shrink the size of the subsample (i.e., a higher J could only lead to a smaller $|A'|$), it would allow for the inclusion of more confounding factors in the model. Of particular interest would be balancing by subjects' geographic location (though, if this were encoded as a continuous variable, it would require a more sophisticated S), as well as volumetric measurements of finer areas of the brain (e.g., for a functional connectivity study, balancing by hemispheric or amygdala volume).

Since the experiments in Chapters 7 and 8 were undertaken, some of these problems have since been addressed with three important innovations in the class balancing algorithm. These are: (1) a recursive method by which the class balancing algorithm is repeatedly applied on excluded data, then added back, as a means of maximizing the count of data included; (2) different means of discretizing continuous covariates – notably, averaging between the discretization values found by number of datapoints and those found with equidistant points between the minimum and maximum of all continuous values; and (3) generalizing the algorithm to multiple, rather than two, classes, avoiding the need to apply the algorithm twice in quartile balancing.

A potential alternative to class balancing is regressing confounding factors from datapoints directly. Regressing covariates from individual MRI data has been the subject of extensive research, especially with regards to the effects of head movement (Kundu et al., 2013; Patel and Bullmore, 2016), though properly modelling spin-echo effects that lasts several seconds

after head movement is particularly challenging. In small datasets where removing any individual datapoint may adversely affect the statistical strength of the study, such regression is necessary. Nonetheless, with data as high-dimensional and complex as MRI, such confounding factors are incredibly difficult to properly regress from individual datapoints, to the extent that it is not detectable by a deep learning model. By comparison, class balancing is trivial.

An interesting future use of deep learning may be to regress out these confounding factors from individual datasets with the use of adversarial neural networks (Goodfellow et al., 2014), which could be tasked with regressing out confounding factors (i.e., making it such that a measurement such as motion were undetectable by a deep learning model, while altering the data as minimally as possible). However, this is an entire research project in itself.

9.5 Comparison of machine learning encoding methods

This thesis presented the use of two different methods of encoding brain connectivity data in convolutional neural networks: vertical convolutions (Chapter 4) and stochastic convolutions (Chapters 7 and 8). This represented a transition from encoding edges in a network by their natural groupings (i.e., edges they are attached to), which has a sounder theoretical basis, to a method that grouped them randomly. However, in addition to outputting class activation maps that did not have a spatial bias, these stochastic convolutions produced either similar or higher classification accuracies than the vertical convolutions. While it is possible that other factors may have played a role in these results, they nonetheless indicate that such encoding may be an unnecessary prior.

These results were surprising, because it is established that analogous priors are very necessary in photographic image classifications, from which the present methods are directly derived (Krizhevsky et al., 2012). CNNs designed to classify such images employ square-shaped convolutional filters, for two primary reasons. First, these convolutions encode information about the relative spatial organization of features. Square-shaped convolutional filters group pixels that are spatially close together, carrying the inherent (and well-founded) assumption that adjacent pixels encode more useful information when considered as a group than as independent components (which would be the case with fully-connected neural network). Second, by repeatedly convolving many input features into a single output value, they act as a means of regularization that prevent overfitting.

I offer possible explanations for why each of these effects, which benefits CNNs for 2D image classification, may not benefit brain connectome classification using vertical or cross-shaped convolutions in exactly the same way, offering a reason for the superiority of stochastic encoding. First, grouping edges together by nodes, while intuitive and simple, is likely not the most ideal means of grouping edges together in the first place, since functional brain networks are more often organized as networks distributed over a wide area. This prior assumes that connections to common nodes matter more than other factors, such as spatial positions or symmetric similarity; for instance, an edge strongly connecting the left amygdala to the left hippocampus is more likely to have an interesting association with the edge connecting the right amygdala to the right hippocampus, than to an edge connecting the left amygdala to the right cerebellum. Another issue with encoding by nodes is scalability: more values are convolved if the size of the adjacency matrix is scaled up (which can happen as the result of a finer parcellation), which dilutes each individual values' contribution. While different parcellation sizes were not tested in this thesis, it is a theoretical issue worth noting.

Second, the regularization effect, while still preventing overfitting, could be counterproductive for connectomes. In effect, convolutions compress information into a single value; each vertical convolution, for an input of edges, $e_1, e_2, ...e_{116}$, is a function that compresses these edges into a single value, $C = e_1w_1 + e_2w_2... + e_{116}w_{116}$. However, as shown in the various class activation maps displayed throughout this thesis, each edge does not contribute equally to the output decision of the deep learning model, as the values of some edges hold more influence on the output decision than others. Furthermore, because different areas of the brain play different roles in classification, these "important" edges are not distributed evenly across nodes. Thus, grouping edges by node would inevitably compress a disproportionate number of edges with higher influence on the output classification, thus contributing to an unnecessary bottleneck of important information flowing through the network. Conversely, in 2D image classification, it is usually not the case that fine variations in the values of individual pixels have a large effect on the outcome, but rather the location and general shape of objects detected. Stochastic encoding for brain networks would more evenly distribute important edges across each convolution used and minimize any such informational bottlenecks, while still helping to prevent overfitting.

Notably, however, while stochastic models do away with spatial encoding, they maintain *depth*, that is, the pairing of multiple measurements of an edge (multiple wavelet frequencies, or functional and structural connectivities) in multi slice connectomes, which was essential for studying the roles of different layers of the input connectome in a classification task.

A future direction of interest would be to experiment with different ways of encoding graphs in CNNs that use more sophisticated priors. Indeed, this is a potent potential application of the work of Chapter 3: pathways of interest could be encoded and compared directly in a deep learning model, much in the same way that objects are encoded in CNNs for 2D image classification. A practical constraint, however, is that one has to work with the machine learning libraries developed within pure computer science, which may not be easily generalizable to new forms of data found in different fields. This is discussed further in Section 9.8, which notes my previous efforts in this direction, and Section 9.9, which discusses limitations in current deep learning frameworks for encoding brain networks.

9.6 Interpretation of visualization methods

Three deep learning visualization methods were used in this thesis: activation maximization, occlusion, and class activation mapping. The first of these was used to detect ways in which the deep learning model clustered datasets, and so it is not directly comparable to the other two methods, which both have outputs of the same dimension as the input data that indicate which parts of an input dataset were salient. However, while both indicate salience, several considerations must be made in their interpretation.

In their original conception, neither class activation mapping nor occlusion were used in a quantitative way (Zeiler and Fergus, 2013; Simonyan et al., 2014; Selvaraju et al., 2017), but rather as visual aids that informally validated the correctness of photographic classification algorithms, so their quantification in the context of other scientific data deserves further scrutiny. In Chapter 7, it was found that class activation map values actually did not vary substantially by dataset input; this is perhaps because, unlike 2D images, connectivity input data, having already been registered to the same space, has no spatial variation, and so the model and visualization method simply focus on the same areas for every input dataset. Thus, this method did show which areas were most crucial in classification, but not how they were used, or for which subgroups.

This was not an issue, however, for occlusion outputs, which did show the ability to vary across groups, shedding more light on which areas (in the context of Chapter 7, brain networks) were most crucial in phenotypic characterization. However, occlusion bore computing power limits substantially higher than class activation maps required; moreover, most occlusion applications in pure deep learning were applied to a single input dataset (i.e., one

photo) with one CNN, and because I sought occlusion-based salience maps on several thousand datasets, rather than a single dataset, on an ensemble of CNNs, many occlusion-based methods from pure deep learning bore a prohibitively high computational cost for the purposes of this work.

In summary, class activation mapping indicates which parts input data were most important for classification, without indicating how or for which subgroups. Occlusion is capable of this fine-tuned analysis, but computational costs limit it to testing only a-priori assumptions. However, it is likely that future innovations can improve the applicability to occlusion to brain data.

9.7 Shortcomings of salience detection methods

Class activation mapping and occlusion are both salience detection methods borrowed from photographic image classification. In that context, the ways in which salient areas differ or aid in classification may be self-evident to human interactors, but this is not the case for connectivity. A major shortcoming of both these methods is that they do not indicate whether a stronger or weaker connectivity between edges and networks drove the classification, or whether it was driven by a complex interaction between different connectivity values that would require more sophisticated means to characterize. It is likely that further work on these methods would be required in order to enrich the explanations of salience detection in brain connectivity.

9.8 Failed research directions

There were a number of experiments which, though they initially seemed promising, failed to yield any results after several lines of inquiry. These are noted here.

In an attempt to add meaningful priors to the CNN stochastic encoding method, different means of scrambling edge orders were experimented with. For instance, edges that were spatially closer to one another were placed in the same rows, and edges belonging to the same pathway were placed in the same row. However, multiple attempts at such encoding methods yielded no discernible improvements in accuracy over purely random orderings. Additionally, a single random ordering was tested over multiple training/test/validation set

splits, but its distribution of AUROCs was not found to be any different from the use of multiple random orderings, leading to the conclusion that no one random ordering bears any notable advantages in accuracy over others. In general, lines of exploration that attempted to find more favourable patterns in different stochastic orderings of edges ended in failure.

Classification tasks other than those noted in this thesis were attempted, specifically on the functional connectivity of subclinical depression patients in the UK Biobank (of which there were around five thousand, making it an ideal covariate to study in a big data context) and patients with auditory or visual hallucinations in the BioBank. While a study of depression would have not only been of general scientific interest, but also a means of comparing results in Chapter 3 directly to our results, both depression and hallucination classification tests yielded accuracies close to random.

In a contradiction with the normative pathway results in Chapter 3, which found partial derivatives to be an effective means of functional connectivity estimation, this did not yield favorable results when applied to deep learning. Attempts to classify partial correlation matrices usually yielded much worse results than those that used Pearson or wavelet correlation. This is perhaps a testament to contextual importance of different methodologies.

Further experiments with multi-modal methodologies from Chapter 8 were made to encode covariates directly. In particular, layers were added that encoded both age and collection for the classification of autism. However, this was found to not yield any notable differences in model performance over the combined structure/function encoding alone. It is notable that others have had success with encoding covariates directly (Jonsson et al., 2019), so possible explanations for this may be that the encoding method used was wrong, or simply that the deep learning model inferred the information already from the data.

Several alternative neural network frameworks were researched and attempted before the convolutional neural network framework was settled on. These include variations of node wise classification graphs (Bruna et al., 2014; Defferrard et al., 2016; Kipf and Welling, 2017); models that turned graphs into 2D images and sent them through a 2D image CNN (Hechtlinger et al., 2017); adaptation of node wise graph classification models, in which whole-graph classification could theoretically be achieved by maintaining the same label across a whole graph (Hamilton et al., 2017); and a number of models that simply derived unrelated global or nodewise measures from a graph and considered these independently (Nikolentzos et al., 2017). These experiments failed or were deemed unsuitable for a number of reasons, ranging from the purely theoretical to poor software design and documentation.

9.9 Future directions

In this section, I discuss potential future directions to build on and improve the present body of work.

Many improvements to the methods presented in this thesis may be found in *scale*, which this thesis leaves plenty of room for. For instance, increasing the number of parcellations used, the number of input channels, the depth and breadth of connectomes, the size of the training dataset, or the use of techniques that require more computing power (either on the preprocessing side, which would likely improve the quality of the registrations, or the machine learning side, allowing for the use of more advanced models) – all would have a strong chance of increasing the accuracy of the results.

Additionally, improvements in preprocessing techniques in general may offer up to an approximate 50% increase in datapoints, as the fully-automated techniques used in this thesis had very high failure rates and required the elimination of an extraordinary amount of data from consideration. Additionally, as explained in Chapter 2, computational limitations prevented thorough grid searches for hyperparameter tuning, and in general those are likely to offer at least modest improvements to classification accuracy. Early on in the research, some informal grid searching was undertaken to improve hyperparameters for models, but a similar search was not undertaken on the pre-processing techniques used (for instance, comparisons of different parcellations), in which I simply used the best conventional methods available. An increase in computing time and power could be utilized in many ways for the benefit of machine learning on big data in MRI.

As a long-term research goal, fundamental advancements in deep learning architecture could substantially improve results for the benefit of this research. This work attempted to undertake this to some degree; I eliminated the spatial encoding biases introduced from 2D-image-based deep learning models in order to improve the resolution of class activation maps. However, my models fail to explicitly encode subnetworks and pathways that are likely present in brain connectome data. In fact, as discussed above, my work departed from using any spatial priors: previous attempts to fully encode graphs by edge-to-edge connections using deep learning libraries were undertaken by Kawahara et al. (2017) and replicated in Leming and Suckling (2019), but I later posited that that architecture is flawed, since the cross-shaped convolutions are created by adding independent vertical and horizontal filters, and the backpropagation algorithm is not necessarily optimized to deal with convolutional

filters being added together in this way; switching from cross-shaped to vertical filters, in the form presented in Chapter 4, not only eliminated recurring problems of exploding/imploding gradients present in the cross-shaped filters, but increased accuracy generally.

A basis of this problem is that modern deep learning implementations were specifically designed and optimized for 2D images and, later, natural language understanding, and scientists in non-computing-centered fields attempting to benefit from them must adapt these models to their own data, rather than design models with priors befitting their own data formats. Unfortunately, this not only requires extraordinary computational expertise, but also extraordinary understanding of data formats in unrelated scientific fields. Pioneers in computer vision benefit not only from greater understanding of their novel machine learning techniques, but also greater understanding in the types of data they work with (i.e., the average human interactor intuitively understands photographic images and natural language more than a neuroscientist understands MRIs).

Practically speaking, raw MRIs differ from photographic images because they are three- or four-dimensional, spatially invariant, and contain many features which the machine learning algorithm would preferably ignore, such as motion and head size (though this thesis has gone to great lengths to address the last issue). Moreover, it is more often individual objects that are of interest in photographic images, while in MRIs, just as often, interesting information lies in the relationship of disparate objects; this is the heart of the interest in functional connectivity and structural covariance. In 2D image recognition, this is analogous to "scene understanding", which is a far more difficult task than individual object classification, and for which simple convolutional neural networks are not necessarily equipped to evaluate, though much progress has been made on it in recent years.

Improvements in connectomic classification may come from graph-theory-based machine learning models, which is currently an active area of research, but for which most interest is presently in areas such as social networks — graphs that are different in form than brain networks (i.e., without fixed nodes). Nonetheless, given interest in the field in general, it is likely than an applicable model, capable of encoding pathways and communities explicitly, will be produced in the near future.

9.10 Conclusion

I have presented, in this thesis, a comprehensive framework for classifying large amounts of MRI brain connectivity data, using novel methods on a large, accumulated dataset. I have also presented methods adapted from computer science to analyze these brain connectomes, providing novel insights into which features drive phenotypic differences across populations.

Acknowledgements

The study was funded by the UK Medical Research Council (grant: G0802226), the National Institute for Health Research (NIHR) (grant: 06-05-01), financial support from the Department of Health, and the Behavioral and Clinical Neuroscience Institute (BCNI), University of Cambridge, the latter being jointly funded by the Medical Research Council and the Wellcome Trust. Additional support was received from the Cambridge Biomedical Research Centre. Matthew Leming was funded by a Gates Cambridge Scholarship from the University of Cambridge. We also thank the support from Alzheimer's Research UK (ARUK-SRF2017B-1).

Special thanks go to all participants for their contribution to this work.We also greatly appreciate the role of the Wolfson Brain Imaging Centre, Cambridgeshire and Peterborough NHS Foundation Trust, Child and Adolescent Mental Health Services, Mental Health Research Network, IMPACT research assistants, and IMPACT clinicians, without whom this study could not have taken place.

This study used publicly available datasets, each with their own acknowledgements. We recognise the contributions of the Alzheimer's Disease Neuroimaging Initiative, International Consortium for Brain Mapping, National Database for Autism Research, NIH Pediatric MRI Data Repository, National Database for Clinical Trials, Research Domain Criteria Database, Adolescent Brain Cognitive Development Study, UK Biobank Resource, 1000 Functional Connectomes Project, ABIDE I and II, and Open fMRI. This research was co-funded by the NIHR Cambridge Biomedical Research Centre and Marmaduke Sheild. Matthew Leming is supported by a Gates Cambridge Scholarship from the University of Cambridge.

ADNI Acknowledgement

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD

ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

ICBM Acknowledgement

Data collection and sharing for this project was provided by the International Consortium for Brain Mapping (ICBM; Principal Investigator: John Mazziotta, MD, PhD). ICBM funding was provided by the National Institute of Biomedical Imaging and BioEngineering. ICBM data are disseminated by the Laboratory of Neuro Imaging at the University of Southern California.

NIMH Database Acknowledgements

NDAR Acknowledgement Data and/or research tools used in the preparation of this manuscript were obtained from the NIHsupported National Database for Autism Research (NDAR). NDAR is a collaborative informatics system created by the National Institutes of Health to provide a national resource to support and accelerate research in autism. Dataset identifier(s): [NIMH Data Archive Collection ID(s) or NIMH Data Archive Digital Object Identifier (DOI)]. This manuscript reflects the views of the authors and may not reflect the opinions or views of the NIH or of the Submitters submitting original data to NDAR.

Pediatric MRI Acknowledgement Data used in the preparation of this article were obtained from the NIH Pediatric MRI Data Repository created by the NIH MRI Study of Normal Brain Development. This is a multisite, longitudinal study of typically developing children from ages newborn through young adulthood conducted by the Brain Development Cooperative Group and supported by the National Institute of Child Health and Human Development, the National Institute on Drug Abuse, the National Institute of Mental Health, and the National Institute of Neurological Disorders and Stroke (Contract #s N01-HD02-3343, N01-MH9-0002, and N01-NS-9-2314, -2315, -2316, -2317, -2319 and -2320). A listing of the participating sites and a complete listing of the study investigators can be found at http://pediatricmri.nih.gov/nihpd/info/participating_centers.html. Dataset identifier(s): [NIMH Data Archive Collection ID(s) or NIMH Data Archive Digital Object Identifier (DOI)].

NDCT Acknowledgement

Data and/or research tools used in the preparation of this manuscript were obtained and analyzed from the controlled access datasets distributed from the NIMH-supported National Database for Clinical Trials (NDCT). NDCT is a collaborative informatics system created by the National Institute of Mental Health to provide a national resource to support and accelerate discovery related to clinical trial research in mental health. Dataset identifier(s): [NIMH Data Archive Collection ID(s) or NIMH Data Archive Digital Object Identifier (DOI)].

RDoCdb Acknowledgement

Data and/or research tools used in the preparation of this manuscript were obtained and analyzed from the controlled access datasets distributed from the NIMH-supported Research Domain Criteria Database (RDoCdb). RDoCdb is a collaborative informatics system created by the National Institute of Mental Health to store and share data resulting from grants funded through the Research Domain Criteria (RDoC) project. Dataset identifier(s): [NIMH Data Archive Collection ID(s) or NIMH Data Archive Digital Object Identifier (DOI)].

ABCD Acknowledgement

Data used in the preparation of this article were obtained from the Adolescent Brain Cognitive Development (ABCD) Study (https://abcdstudy.org), held in the NIMH Data Archive (NDA). This is a multisite, longitudinal study designed to recruit more than 10,000 children age 9-10 and follow them over 10 years into early adulthood. The ABCD Study is supported by the National Institutes of Health and additional federal partners under award numbers

U01DA041022, U01DA041028, U01DA041048, U01DA041089, U01DA041106, U01DA041117, U01DA041120, U01DA041134, U01DA041148, U01DA041156, U01DA041174, U24DA041123, and U24DA041147. A full list of supporters is available at https://abcdstudy.org/federal-partners.html. A listing of participating sites and a complete listing of the study investigators can be found at https://abcdstudy.org/Consortium_Members.pdf. ABCD consortium investigators designed and implemented the study and/or provided data but did not necessarily participate in analysis or writing of this report. This manuscript reflects the views of the authors and may not reflect the opinions or views of the NIH or ABCD consortium investigators.

UK Biobank Acknowledgement

This research has been conducted using the UK Biobank Resource [project ID 20904]. This research was co-funded by the NIHR Cambridge Biomedical Research Centre and a Marmaduke Sheild grant to Richard A.I. Bethlehem and Varun Warrier. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

Other database Aknowledgements

We would also like to thank the 1000 Functional Connectomes Project, ABIDE I and II, and Open fMRI.

Bibliography

Abdelnour, F., Voss, H., and Raj, A. (2014). Network diffusion accurately models the relationship between structural and functional brain connectivity networks. *NeuroImage*, 90:335–347.

Abraham, A., Milham, M., Di Martino, A., Craddock, R., Samaras, D., Thirion, B., and Varoquaux, G. (2016). Deriving reproducible biomarkers from multi-site resting-state data: An Autism-based example. *NeuroImage*, 147:736–745.

Achard, S. and Bullmore, E. (2007). Efficiency and cost of economical brain functional networks. *PLoS Comput Biol*, 3:e17.

Achard, S., Salvador, R., Whitcher, B., Suckling, J., and Ed Bullmore, E. (2006). A resilient, low-frequency, small-world human brain functional network with highly connected association cortical hubs. *Journal of Neuroscience*, 26:63–72.

Acharya, U., Oh, S., Hagiwara, Y., Tan, J., , and Adeli, H. (2018a). Deep convolutional neural network for the automated detection of seizure using EEG signals. *Computers in Biology and Medicine*, 100:270–278.

Acharya, U., Oh, S., Hagiwara, Y., Tan, J., Adeli, H., and Subha, D. (2018b). Automated EEG-based Screening of Depression Using Deep Convolutional Neural Network. *Computer Methods and Programs in Biomedicine*, 161:103–113.

Adachi, Y., Osada, T., Sporns, O., Watanabe, T., Matsui, T., Miyamoto, K., and Miyashita, Y. (2012). Functional connectivity between anatomically unconnected areas is shaped by collective network-level effects in the macaque cortex. *Cereb Cortex*, 22:1586–92.

Adenauer, H., Pinosch, S., Catani, C., Gola, H., Keil, J., Kissler, J., and Neuner, F. (2010). Early processing of threat cues in posttraumatic stress disorder-evidence for a cortical vigilance-avoidance reaction. *Biological Psychiatry*, 68:451–458.

Agcaoglu, O., Miller, R., Mayer, A., Hugdahl, K., and Calhoun, V. (2015). Lateralization of resting state networks and relationship to age and gender. *Neuroimage*, 104:310–325.

Ahmadlou, M., Adeli, H., and Adeli, A. (2010). Fractality and a Wavelet-Chaos-Neural Network Methodology for EEG-based Diagnosis of Autistic Spectrum Disorder. *Journal of Clinical Neurophysiology*, 27:328–333.

Ahmadlou, M., Adeli, H., and Adeli, A. (2012). Fuzzy Synchronization Likelihood-Wavelet Methodology for Diagnosis of Autism Spectrum Disorder. *Journal of Neuroscience Methods*, 211:203–209.

Akkus, Z., Galimzianova, A., Hoogi, A., Rubin, D., and Erickson, B. (2017). Deep Learning for Brain MRI Segmentation: State of the Art and Future Directions. *J Digit Imaging.*, 30:449–459.

Al-Zubaidi, A., Mertins, A., Heldmann, M., Jauch-Chara, K., and M unte1, T. (2019). Machine Learning Based Classification of Resting-State fMRI Features Exemplified by Metabolic State (Hunger/Satiety). *Front. Hum. Neurosci.*, 13.

Alarcón, G., Pfeifer, J., Fair, D., and Nagel, B. (2018). Adolescent Gender Differences in Cognitive Control Performance and Functional Connectivity Between Default Mode and Fronto-Parietal Networks Within a Self-Referential Context. *Front Behav Neurosci.*, 12:17.

Aleman, A., Kahn, R., and Selten, J. (2003). Sex differences in the risk of schizophrenia: evidence from meta-analysis. *Arch Gen Psychiatry*, 60:565–571.

Aleman, A. and Swart, M. (2008). Sex differences in neural activation to facial expressions denoting contempt and disgust. *PloS One*, 3:e3622.

Alexander-Bloch, A., Giedd, J., and Bullmore, E. (2013a). Imaging structural co-variance between human brain regions. *Nat. Rev. Neurosci.*, 14:322–336.

Alexander-Bloch, A., Vértes, P., Stidd, R., Lalonde, F., Clasen, L., Rapoport, J., Giedd, J., Bullmore, E., and Gogtay, N. (2013b). The anatomical distance of functional connections predicts brain network topology in health and schizophrenia. *Cereb. Cortex*, 23:127–138.

Allen, E., Damaraju, E., Plis, S., Erhardt, E., Eichele, T., and Calhoun, V. (2014). Tracking whole-brain connectivity dynamics in the resting state. *Cereb Cortex*, 24:663–676.

Allen, E., Erhardt, E., Wei, Y., Eichele, T., and Calhoun, V. (2012). Capturing inter-subject variability with group independent component analysis of fMRI data: a simulation study. *NeuroImage*, 59:4141–4159.

Alstott, J., Breakspear, M., Hagmann, P., Cammoun, L., and Sporns, O. (2009). Modeling the impact of lesions in the human brain. *PLoS Comput. Biol.*, 5:e1000408.

Amunts, K., Malikovic, A., Mohlberg, H., Schormann, T., and Zilles, K. (2000). Brodmann's areas 17 and 18 brought into stereotaxic space-where and how variable? *NeuroImage*, 11:66–84.

Anderson, A. and Thomason, M. (2013). Functional plasticity before the cradle: a review of neural functional imaging in the human fetus. *Neurosci Biobehav Rev.*, 37:2220–2232.

Anderson, J., Druzgal, T., Froehlich, A., DuBray, M., Lange, N., Alexander, A., Abildskov, T., Nielsen, J., Cariello, A., Cooperrider, J., Bigler, E., and Lainhart, J. (2011a). Decreased interhemispheric functional connectivity in autism. *Cereb Cortex*, 21:1134–1146.

Anderson, J., Nielsen, J., Froehlich, A., DuBray, M., Druzgal, T., Cariello, A., Cooperrider, J., Zielinski, B., Ravichandran, C., Fletcher, P., Alexander, A., Bigler, E., Lange, N., and Lainhart, J. (2011b). Functional connectivity magnetic resonance imaging classification of autism. *Brain*, 134:3742–3754.

Anderson, N., Harenski, K., Harenski, C., Koenigs, M., Decety, J., Calhoun, V., and Kiehl, K. (2019). Machine learning of brain gray matter differentiates sex in a large forensic sample. *Human Brain Mapp.*, 40:1496–1506.

Anderson, V., Catroppa, C., Morse, S., Haritou, F., and Rosenfeld, J. (2005). Functional plasticity or vulnerability after early brain injury? *Pediatrics*, 116:1374–1382.

Andreano, J., Dickerson, B., and Barrett, L. (2014). Sex differences in the persistence of the amygdala response to negative material. *Soc Cogn Affect Neurosci*, 9:1388–1394.

Ansari, A., Cherian, P., Caicedo, A., Naulaers, G., De Vos, M., and Van Huffel, S. (2019). Neonatal Seizure Detection Using Deep Convolutional Neural Networks. *International Journal of Neural Systems*, 29:1850011.

Antoniades, A., Spyrou, L., Martin-Lopez, D., Valentin, A., Alarcon, G., Sanei, S., and Took, C. (2018). Deep neural architectures for mapping scalp to intracranial EEG. *International Journal of Neural Systems*, 28.

Arbabshirani, M., Plis, S., Sui, J., and Calhoun, V. (2017). Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *NeuroImage*, 145(Pt B):137–165.

Archer, J. (2004). Sex Differences in Aggression in Real-World Settings: A Meta-Analytic Review. *Review of General Psychology*, 8:291–322.

Arnett, A., Pennington, B., Peterson, R., Willcutt, E., DeFries, J., and Olson, R. (2017). Explaining the sex difference in dyslexia. *J Child Psychol Psychiatry*, 58:719–727.

Ashburner, J. and Friston, K. (2000). Voxel-based morphometry: the methods. *Neuroimage*, 11:805–821.

Assaf, M., Jagannathan, K., Calhoun, V., Miller, L., Stevens, M., Sahl, R., O'Boyle, J., Schultz, R., and Pearlson, G. (2010). Abnormal functional connectivity of default mode sub-networks in autism spectrum disorder patients. *NeuroImage*, 53:247–256.

Avena-Koenigsberger, A., Misic, B., Hawkins, R., Griff, A., Hagmann, P., Goni, J., and Sporns, O. (2017). Path ensembles and a tradeoff between communication efficiency and resilience in the human connectome. *Brain Struct Funct*, 222:603–618.

Baibakov, S. and Fedorov, V. (2010). Morphometric characteristics of the brain in children aged one year (magnetic resonance tomography data). *Neurosci. Behav. Physiol.*, 40:69–72.

Balardin, J., Comfort, W., Daly, E., Murphy, C., Andrews, D., Murphy, D., Ecker, C., Consortium, M. A., and Sato, J. (2015). Decreased centrality of cortical volume covariance networks in autism spectrum disorders. *J Psychiatr Res.*, 69:142–149.

Barch, D., Burgess, G., Harms, M., Petersen, S., Schlaggar, B., Corbetta, M., Glasser, M., Curtiss, S., Dixit, S., Feldt, C., Nolan, D., Bryant, E., Hartley, T., Footer, O., Bjork, J., Poldrack, R., Smith, S., Johansen-Berg, H., Snyder, A., Van Essen, D., and Consortium., W.-M. H. (2013). Function in the human connectome: task-fMRI and individual differences in behavior. *Neuroimage*, 80:169–189.

Baron-Cohen, S. (1988a). An assessment of violence in a young man with Asperger's syndrome. *JChild Psychol Psychiatry*, 29:351–360.

Baron-Cohen, S. (1988b). Social and pragmatic deficits in autism: cognitive or affective? *J Autism Dev Disord*, 18:379–402.

Baron-Cohen, S. (1988c). Without a theory of mind one cannot participate in a conversation. *Cognition*, 29:83–84.

Baron-Cohen, S. (2004). The cognitive neuroscience of autism. *J Neurol Neurosurg Psychiatry*, 75:945–948.

Baron-Cohen, S., Lombardo, M., Auyeung, B., Ashwin, E., Chakrabarti, B., and Knickmeyer, R. (2011). Why Are Autism Spectrum Conditions More Prevalent in Males? *PLoS Biol*, 9:e1001081.

Baron-Cohen, S., Ring, H., Moriarty, J., Schmitz, B., Costa, D., and Ell, P. (1994). Recognition of mental state terms – clinical findings in children with autism and a functional neuroimaging study of normal adults. *Br J Psychiatry*, 165:640–649.

Barttfeld, P., Wicker, B., Cukier, S., Navarta, S., Lew, S., Leiguarda, R., and Sigman, M. (2012). State-dependent changes of connectivity patterns and functional brain network topology in autism spectrum disorder. *Neuropsychologia*, 50:3653–3662.

Basser, P., Mattiello, J., and LeBihan, D. (1994). Estimation of the effective self-diffusion tensor from the NMR spin echo. *J. Magn. Reson. B*, 103:247–254.

Bassett, D. and Bullmore, E. (2006). Small-World Brain Networks. *The Neuroscientist*, 12:512–523.

Bassett, D., Porter, M., Wymbs, N., Grafton, S., Carlson, J., and Mucha, P. (2013). Robust detection of dynamic community structure in networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 23:013142.

Bassett, D., Wymbs, N., Porter, M., Mucha, P., Carlson, J., and Grafton, S. (2011). Dynamic reconfiguration of human brain networks during learning. *PNAS*, 108:7641–7646.

Bastos, A. and Schoffelen, J. (2016). A Tutorial Review of Functional Connectivity Analysis Methods and Their Interpretational Pitfalls. *Frontiers in Systems Neuroscience*, 9.

Batista-García-Ramó, K. and Ivette Fernández-Verdecia, C. (2018). What We Know About the Brain Structure–Function Relationship. *Behav Sci (Basel)*, 8:39.

Bavelas, A. (1950). Communication patterns in task-oriented groups. *J. Acoust. Soc. Am*, 22:725–730.

Behrens, T., Berg, H., Jbabdi, S., Rushworth, M., and Woolrich, M. (2007). Probabilistic diffusion tractography with multiple fibre orientations: What can we gain? *NeuroImage*, 34:144–155.

Ben Bashat, D., Kronfeld-Duenias, V., Zachor, D., Ekstein, P., Hendler, T., Tarrasch, R., Even, A., Levy, Y., and Ben Sira, L. (2007). Accelerated maturation of white matter in young children with autism: A high b value DWI study. *NeuroImage*, 37:40–47.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57:289–300.

Bergamo, A., Bazzani, L., Anguelov, D., and Torresani, L. (2014). Self-taught object localization with deep networks. *IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Berlingerio, M., Coscia, M., Giannotti, F., Monreale, A., and Pedreschi, D. (2011). Foundations of Multidimensional Network Analysis. *International Conference on Advances in Social Networks Analysis and Mining*.

Betzel, R., Byrge, L., He, Y., Goñi, J., Zuo, X., and Sporns, O. (2014). Changes in structural and functional connectivity among resting-state networks across the human lifespan. *NeuroImage*, 102:345–357.

Betzel, R., Fukushima, M., He, Y., Zuo, X., and Sporns, O. (2016). Dynamic fluctuations coincide with periods of high and low modularity in resting-state functional brain networks. *NeuroImage*, 127:287–297.

Bhat, S., Acharya, U., Adeli, Muralidhar Bairy, G., and Adeli, A. (2014a). Autism: Cause Factors, Early Diagnosis and Therapies. *Reviews in the Neurosciences*, 25:841–850.

Bhat, S., Acharya, U., Adeli, H., Muralidhar Bairy, G., and Adeli, A. (2014b). Automated Diagnosis of Autism: In Search of a Mathematical Marker. *Reviews in the Neurosciences*, 25:851–861.

Biederman, J. (2005). Attention-deficit/hyperactivity disorder: a selective overview. *Biol. Psychiatry*, 57:1215–1220.

Bijsterbosch, J., Woolrich, M., Glasser, M., Robinson, E., Beckmann, C., Van Essen, D., Harrison, S., and Smith, S. (2018). The relationship between spatial configuration and functional connectivity of brain regions. *eLife*, 7:e32992.

Bitsko, R., Holbrook, J., Visser, S., Mink, J., Zinner, S., Ghandour, R., and Blumberg, S. (2014). A national profile of Tourette syndrome, 2011-2012. *J Dev Behav Pediatr*, 35:317–322.

Bloch, F., Hansen, W., and Packard, M. (1946). Nuclear Induction. *Phys. Rev.*, 69.

Bluhm, R., Osuch, E., Lanius, R., Boksman, K., Neufeld, R., Théberge, J., and Williamson, P. (2008). Default mode network connectivity: effects of age, sex, and analytic approach. *Neuroreport.*, 19:887–891.

Boguña, M., Krioukov, D., and Claffy, K. (2009). Navigability of complex networks. *Nat Phys*, 5:74–80.

Bora, E., Harrison, B., Y ucel, M., and Pantelis, C. (2013). Cognitive Impairment in Euthymic Major Depressive Disorder. *Psychol Med.*, 43:2017–2026.

Borgatti, S. (2005). Centrality and network flow. *Social Networks*, 27:55–71.

Borsboom, D. and Cramer, A. (2013). Network Analysis: An Integrative Approach to the Structure of Psychopathology. *Annu. Rev. Clin. Psychol.*, 9:91–121.

Brambilla, P., Hardan, A., di Nemi, S., Perez, J., Soares, J., and Barale, F. (2003). Brain anatomy and development in autism: review of structural MRI studies. *Brain Res Bull*, 61:557–569.

Bressler, S. and Menon, V. (2010). Large-scale Brain Networks in Cognition: Emerging Methods and Principles. *Trends Cogn Sci*, 14:277–290.

Brookhart, M., Schneeweiss, S., Rothman, K., Glynn, R., Avorn, J., and Stürmer, T. (2006). Variable selection for propensity score models. *Am J Epidemiol*, 163:1149–1156.

Brown, C. and Hamarneh, G. (2016). Machine Learning on Human Connectome Data from MRI. *arXiv*.

Brown, C., Kawahara, J., and Hamarneh, G. (2018). Connectome Priors in Deep Neural Networks to Predict Autism. *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*.

Br uhl, A., Rufer, M., Delsignore, A., Kaffenberger, T., Jancke, L., and Herwig, U. (2011). Neural correlates of altered general emotion processing in social anxiety disorder. *Brain Research*, 1378:72–83.

Bruna, J., Zaremba, W., Szlam, A., and LeCun, Y. (2014). Spectral networks and locally connected networks on graphs. *ICLR*.

Bryson, S., Zwaigenbaum, L., Brian, J., Roberts, W., Szatmari, P., Rombough, V., and McDermott, C. (2007). A prospective case series of high-risk infants who developed autism. *J Autism Dev Disord.*, 37:12–24.

Buckner, R., Krienen, F., and Yeo, B. (2013). Opportunities and limitations of intrinsic functional connectivity MRI. *Nat. Neurosci.*, 16:832–837.

Buckner, R., Sepulcre, J., Talukdar, T., Krienen, F., Liu, H., Hedden, T. andrews-Hanna, J., Sperling, R., and Johnson, K. (2009). Cortical hubs revealed by intrinsic functional connectivity: mapping, assessment of stability and relation to Alzheimer's disease. *J. Neurosci.*, 29:1860–1873.

Bullmore, E., Fadili, M., Maxim, V., Sendur, L., Whitcher, B., Suckling, J., Brammer, M., and Breakspear, M. (2004). Wavelets and functional magnetic resonance imaging of the human brain. *NeuroImage*, 23:S234–S249.

Button, K., Ioannidis, J., Mokrysz, C., Nosek, B., Flint, J., Robinson, E. J., and Munafo, M. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14:365–376.

Buzsaki, G., Geisler, C., Henze, D., and Wang, X. (2004). Interneuron diversity series: Circuit complexity and axon wiring economy of cortical interneurons. *Trends Neurosci*, 27:186–193.

Bzdok, D., Altman, N., and Krzywinski, M. (2018). Statistics versus machine learning. *Nature Methods*, 15:233–234.

Caballero-Gaudes, C. and Reynolds, R. (2017). Methods for cleaning the BOLD fMRI signal. *NeuroImage*, 154:128–149.

Cabeza, R. and Nyberg, L. (2000). Imaging Cognition II: An Empirical Review of 275 PET and fMRI Studies. Journal of Cognitive Neuroscience. *J. Cognit. Neurosci.*, 12:1–47.

Calhoun, V. and Adali, T. (2016). Time-varying brain connectivity in fMRI data: whole-brain data-driven approaches for capturing and characterizing dynamic states. *IEEE Signal Process. Mag.*, page 52–66.

Calhoun, V., Adali, T., Pearlson, G., and Pekar, J. (2001). A method for making group inferences from functional MRI data using independent component analysis. *Human Brain Mapping*, 14:140–151.

Calhoun, V., Miller, R., Pearlson, G., and Adal, T. (2014). The chronnectome: time-varying connectivity networks as the next frontier in fMRI data discovery. *Neuron*, 84:262–274.

Cao, M., Wang, J., Dai, Z., Cao, X., Jiang, L., Fan, F., Song, X., Xia, M., Shu, N., Dong, Q., Milham, M., Castellanos, F., Zuo, X., and He, Y. (2014). Topological organization of the human brain functional connectome across the lifespan. *Dev Cogn Neurosci.*, 7:76–93.

Cao, X., Liu, Z., Xu, C., Li, J., Gao, Q., Sun, N., Xu, Y., Ren, Y., Yang, C., and Zhang, K. (2012). Disrupted resting-state functional connectivity of the hippocampus in medication-na ive patients with major depressive disorder. *Journal of Affective Disorders*, 141:194–203.

Cardon, G., Hepburn, S., and Rojas, D. (2017). Structural Covariance of Sensory Networks, the Cerebellum, and Amygdala in Autism Spectrum Disorder. *Front. Neurol.*, 27.

Casanova, R., Whitlow, C., Wagner, B., Espeland, M., and Maldjian, J. (2012). Combining Graph and Machine Learning Methods to Analyze Differences in Functional Connectivity Across Sex. *The Open Neuroimaging Journal*, 6:1–9.

Caseras, X., Mataix-Cols, D., An, S., et al. (2007). Sex differences in neural responses to disgusting visual stimuli: implications for disgust-related psychiatric disorders. *Biological Psychiatry*, 62:464–471.

Casey, B. and Dale, A. (2018). The Adolescent Brain Cognitive Development (ABCD) study: Imaging acquisition across 21 sites. *Dev Cog Neuro*, 32:43–54.

Cauda, F., Geda, E., Sacco, K., D'Agata, F., Duca, S., Geminiani, G., and Keller, R. (2011). Grey matter abnormality in autism spectrum disorder: an activation likelihood estimation meta-analysis study. *J Neurol Neurosurg Psychiatry*, 82:1304–1313.

Cerliani, L., Mennes, M., Thomas, R., Martino, A., Thioux, M., and Keysers, C. (2015). Increased functional connectivity between subcortical and cortical resting-state networks in autism spectrum disorder. *JAMA Psychiatry*, 72:767–777.

Chang, C. and Glover, G. (2010). Time-frequency dynamics of resting-state brain connectivity measured with fMRI. *NeuroImage*, 50:81–98.

Chapin, F. (1947). *Experimental designs in sociological research.* Harper, New York, 1st edition.

Chattopadhyay, S., Tait, R., Simas, T., Nieuwenhuizen, A., Hagan, C., Holt, R., Graham, J., Sahakian, B., Wilkinson, P., Goodyer, I., and Suckling, J. (2017). Cognitive Behavioral Therapy Lowers Elevated Functional Connectivity in Depressed Adolescents. *EBioMedicine*, 17:216–222.

Chen, R., Jiao, Y., and Herskovits, E. (2011). Structural MRI in Autism Spectrum Disorder. *Pediatr Res*, 69:63–68.

Chen, Z., He, Y., Rosa-Neto, P., Germann, J., and Evans, A. (2008). Revealing modular architecture of human brain structural networks by using cortical thickness from MRI. *Cereb. Cortex*, 18:2374–2381.

Cherkassky, V., Kana, R., Keller, T., and Just, M. (2006). Functional connectivity in a baseline resting-state network in autism. *Neuroreport*, 17:1687–90.

Chien, H., Lin, H., Lai, M., Gau, S., and Tseng, W. (2015). Hyperconnectivity of the right posterior temporo-parietal junction predicts social difficulties in boys with autism spectrum disorder. *Autism Res*, 8:427–441.

Cinbis, R., Verbeek, J., and Schmid, C. (2015). Weakly supervised object localization with multi-fold multiple instance learning. *IEEE Trans. on Pattern Analysis and Machine Intelligence*.

Cochran, W. and Rubin, D. (1973). Controlling bias in observational studies: A review. *Sankhya: The Indian Journal of Statistics, Series A.*, 35:417–446.

Cole, M., Yang, G., Murray, J., Repovs, G., and Anticevic, A. (2016). Functional connectivity change as shared signal dynamics. *J Neurosci Methods*, 259:22–39.

Collins, D. (1998). Zijdenbos, A.P. and Kollokian, V. and Sled, J.G. and Kabani, N.J. and Holmes, C.J. and Evans, A.C. *Trans. Med. Imag.*, 17:463–468.

Cook, J., Blakemore, S., and Press, C. (2013). Atypical basic movement kinematics in autism spectrum conditions. *Brain*, 136:2816–2824.

Copeland, R. (2019). Google's 'Project Nightingale' Gathers Personal Health Data on Millions of Americans. *The Wall Street Journal*.

Corbetta, M. and Shulman, G. (2002). Control of Goal-Directed and Stimulus-Driven Attention in the Brain. *Nature Reviews Neuroscience*, 3:201–215.

Costa, P., Terracciano, A., and McCrae, R. (2001). Gender differences in personality traits across cultures: robust and surprising findings. *J Pers Soc Psychol.*, 81:322–331.

da Fontoura Costa, L. and Travieso, G. (2007). Exploring complex networks through random walks. *Phys Rev E*, 75:016102.

Damadian, R. (1971). Tumor detection by nuclear magnetic resonance. *Science*, 171:1151–1153.

Damoiseaux, J., Prater, K., Miller, B., and Greicius, M. (2012). Functional connectivity tracks clinical deterioration in Alzheimer's disease. *Neurobiol. Aging*, 33:e19–e30.

Deco, G., Jirsa, V., and McIntosh, A. (2011). Emerging concepts for the dynamical organization of resting-state activity in the brain. *Nat Rev Neurosci*, 12:43–56.

Deco, G., Jirsa, V., McIntosh, A., Sporns, O., and K otter, R. (2009). Key role of coupling, delay, and noise in resting brain fluctuations. *PNAS*, 106:10302–10307.

Deco, G., Ponce-Alvarez, A., Mantini, D., Romani, G., Hagmann, P., and Corbetta, M. (2013). Resting-state functional connectivity emerges from structurally and dynamically shaped slow linear fluctuations. *J Neurosci*, 33:11239–11252.

Defferrard, M., Bresson, P., and Vandergheynst, X. (2016). Convolutional neural networks on graphs with fast localized spectral filtering. *NIPS*, pages 3844–3852.

Delbeuck, X., van der Linden, M., and Collette, F. (2003). Alzheimer's disease as a disconnection syndrome? *Neuropsychology*, 13:79–92.

Delmonte, S., O'Gallagher, L., Hanlon, E., McGrath, J., and Balsters, J. (2013). Functional and structural connectivity of frontostriatal circuitry in autism spectrum disorder. *Front Hum Neurosci*, 7:430.

DeRamus, T. and Kana, R. (2015). Anatomical likelihood estimation meta-analysis of grey and white matter anomalies in autism spectrum disorders Author links open overlay panel. *NeuroImage: Clinical*, 7:525–536.

Dey, S., Rao, A., and Shah, M. (2014). Attributed graph distance measure for automatic detection of attention deficit hyperactive disordered subjects. *Front. Neural Circuits*, 8.

Dhillon, P., Wolk, D., Das, S., Ungar, L., Gee, J., and Avants, B. (2014). Subject-specific functional parcellation via Prior Based Eigenanatomy. *NeuroImage*, 99:14–27.

Di Martino, A., Kelly, C., Grzadzinski, R., Zuo, X., Mennes, M., Mairena, M., Lord, C., Castellanos, F., and Milham, M. (2011). Aberrant striatal functional connectivity in children with autism. *Biol Psychiatry*, 69:847–856.

Di Martino, A., O'Connor, D., Chen, B., Alaerts, K. anderson, J., Assaf, M., Balsters, J., Baxter, L., Beggiato, A., Bernaerts, S., Blanken, L., Bookheimer, S., Braden, B., Byrge, L., Castellanos, F., Dapretto, M., Delorme, R., Fair, D., Fishman, I., Fitzgerald, J., Gallagher, L., Keehn, R., Kennedy, D., Lainhart, J., Luna, B., Mostofsky, S., M uller, R., Nebel, M., Nigg, J., O'Hearn, K., Solomon, M., Toro, R., Vaidya, C., Wenderoth, N., White, T., Craddock, R., Lord, C., Leventhal, B., and Milham, M. (2017). Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. *Sci Data*, 4:170010.

Di Martino, A., Yan, C., Li, Q., Denio, E., Castellanos, F., Alaerts, K. anderson, J., Assaf, M., Bookheimer, S., Dapretto, M., Deen, B., Delmonte, S., Dinstein, I., Ertl-Wagner, B., Fair, D., Gallagher, L., Kennedy, D., Keown, C., Keysers, C., Lainhart, J., Lord, C., Luna, B., Menon, V., Minshew, N., Monk, C., Mueller, S., M uller, R., Nebel, M., Nigg, J., O'Hearn, K., Pelphrey, K., Peltier, S., Rudie, J., Sunaert, S., Thioux, M., Tyszka, J., Uddin, L., Verhoeven, J., Wenderoth, N., Wiggins, J., Mostofsky, S., and Milham, M. (2014). The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol Psychiatry*, 19:659–67.

Ding, M., Chen, Y., and Bressler, S. (2006). Granger Causality: Basic Theory and Application to Neuroscience.

Dodonova, Y., Korolev, S., Tkachev, A., and Petrov, D. (2016). Classification of structural brain networks based on information divergence of graph spectra. *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*.

Dolgin, E. (2010). This is your brain online: the Functional Connectomes Project. *Nature Medicine*, 16:351.

Dombi, J. (1982). A general class of fuzzy operators, the DeMorgan class of fuzzy operators and fuzziness measures induced by fuzzy operators. *Fuzzy Sets and Systems*, 8:149–163.

Domes, G., Schulze, L., B ottger, M., et al. (2010). The neural correlates of sex differences in emotional reactivity and emotion regulation. *Human Brain Mapping*, 31:758–769.

Doshi-Velez, F. and Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. *arXiv*.

Du, Y., Fu, Z., and Calhoun, V. (2018). Classification and Prediction of Brain Disorders Using Functional Connectivity: Promising but Challenging. *Front. Neurosci.*, 12:525.

Duncan, N. and Northoff, G. (2013). Overview of potential procedural and participant-related confounds for neuroimaging of the resting state. *J Psychiatry Neurosci.*, 38:84–96.

Eguiluz, V., Chialvo, D., Cecchi, G., Baliki, M., and Apkarian, A. (2005). Scale-free brain functional networks. *Phys Rev Lett*, 94:018102.

Eill, A., Jahedi, A., Gao, Y., Kohli, J., Fong, C., Solders, S., Carper, R., Valafar, F., Bailey, B., and M uller, R. (2019). Functional connectivities are more informative than anatomical variables in diagnostic classification of autism. *Brain Connectivity*, 9.

Elton, A. and Gao, W. (2015). Task-related modulation of functional connectivity variability and its behavioral correlations. *Hum. Brain. Mapp.*, 36:3260–3272.

Emerson, R., Adams, C., Nishino, T., Hazlett, H., Wolff, J., Zwaigenbaum, L., Constantino, J., Shen, M., Swanson, M., Elison, J., Kandala, S., Estes, A., Botteron, K., Collins, L., Dager, S., Evans, A., Gerig, G., Gu, H., McKinstry, R., Paterson, S., Schultz, R., Styner, M., Schlaggar, B., Pruett, J., and Piven, J. (2017). Functional neuroimaging of high-risk 6-month-old infants predicts a diagnosis of autism at 24 months of age. *Sci Transl Med.*, 9:eaag2882.

Eqlimi, E., Riyahi Alam, N., Sahraian, M., Eshaghi, A., Riyahi Alam, S., Ghanaati, H., Firouznia, K., and Karami, E. (2013). Resting State Functional Connectivity Analysis of Multiple Sclerosis and Neuromyelitis Optica Using Graph Theory. *XIII Mediterranean Conference on Medical and Biological Engineering and Computing 2013*, 41:206–209.

Erhan, D., Bengio, Y., Courville, A., and Vincent, P. (2009). Visualizing higher-layer features of a deep network. Technical report 1341, University of Montreal.

Estrada, E. and Hatano, N. (2008). Communicability in complex networks. *Phys Rev E*, 77:036111.

Filippini, N., MacIntosh, B., Hough, M., Goodwin, G., Frisoni, G., Smith, S., Matthews, P., Beckmann, C., and Mackay, C. (2009). Distinct patterns of brain activity in young carriers of the APOE-epsilon4 allele. *PNAS*, 106:7209–7214.

Finn, E., Shen, X., Scheinost, D., Rosenberg, M., Huang, J., Chun, M., Papademetris, X., and Constable, R. (2015). Functional connectome fingerprinting: Identifying individuals using patterns of brain connectivity. *Nat Neurosci.*, 18:1664–1671.

Fischer, H., Sandblom, J., Herlitz, A., Fransson, P., Wright, C., and Backman, L. (2004). Sex-differential brain activation during exposure to female and male faces. *Neuroreport*, 15:235–238.

Fishman, I., Keown, C., Lincoln, A., Pineda, J., and Müller, R. (2014). Atypical cross talk between mentalizing and mirror neuron networks in autism spectrum disorder. *JAMA Psychiatry*, 71:751–760.

Fox, M., Corbetta, M., Snyder, A., Vincent, J., and Raichle, M. (2006). Spontaneous neuronal activity distinguishes human dorsal and ventral attention systems. *PNAS*, 103:10046–10051.

Fox, M., Snyder, A., Vincent, J., Corbetta, M., Essen, D., and Raichle, M. (2005). The human brain is intrinsically organized into dynamic anticorrelated functional networks. *PNAS*, 102:9673–9678.

Fox, M., Zhang, D., Snyder, A., and Raichle, M. (2009). The global signal and observed anticorrelated resting state brain networks. *Journal of Neurophysiology*, 101:3270–3283.

Frau-Pascual, A., Fogarty, M., Fischl, B., Yendiki, A., and Aganj, I. (2019). Quantification of structural brain connectivity via a conductance model. *NeuroImage*, 189:485–496.

Freeman, L. (1977). A Set of Measures of Centrality Based on Betweenness. *Sociometry*, 40:35–41.

Freeman, L. (1979). Centrality in social networks: Conceptual clarification. *Social Networks*, 1:215–239.

Freire, L., Roche, A., and Mangin, J. (2002). What is the Best Similarity Measure for Motion Correction in fMRI Time Series? *IEEE Trans. on Med. Image Anal.*, 21:470–484.

Friston, K. (1994). Functional and Effective Connectivity in Neuroimaging: A Synthesis. *Human Brain Mapping*, 2:56–78.

Friston, K. (2009). Causal Modelling and Brain Connectivity in Functional Magnetic Resonance Imaging. *PLoS Biol*, 7:e1000033.

Friston, K., Frith, C., Little, P., and Frackowiak, R. (1993). Functional Connectivity: The Principal-Component Analysis of Large (PET) Data Sets. *Journal of Cerebral Blood Flow and Metabolism*, 33:5–14.

Frith, U. (1989). A new look at language and communication in Autism. *Br J Disord Commun*, 24:123–150.

Frith, U. (1996). Cognitive explanations of autism. *Acta Paediatr*, 416:63–68.

Fusar-Poli, P., Placentino, A., Carletti, F., Landi, P., Allen, P., Surguladze, S., Benedetti, F., Abbamonte, M., Gasparotti, R., Barale, F., Perez, J., McGuire, P., and Politi, P. (2009). Functional atlas of emotional faces processing: a voxel-based meta-analysis of 105 functional magnetic resonance imaging studies. *J Psychiatry Neurosci*, 34:418–432.

Galán, R. (2008). On how network architecture determines the dominant patterns of spontaneous neural activity. *PLoS One*, 3:e2148.

Gennatas, E., Avants, B., Wolf, D., Satterthwaite, T., Ruparel, K., Ciric, R., Hakonarson, H., Gur, R., and Gur, R. (2017). Age-related effects and sex differences in gray matter density, volume, mass, and cortical thickness from childhood to young adulthood. *J Neurosci.*, 37:5065–5073.

Ghosh, A., Rho, Y., McIntosh, A., K otter, R., and Jirsa, V. (2008). Noise during rest enables the exploration of the brain's dynamic repertoire. *PLoS Comput Biol*, 4:e1000196.

Gibson, E., Li, W., Sudre, C., Fidon, L., Shakir, D., Wang, G., Eaton-Rosen, Z., Gray, R., Doel, T., Hu, Y., Whyntie, T., Nachev, P., Modat, M., Barratt, D., Ourselin, S., Cardoso, M., and Vercauteren, T. (2018). NiftyNet: a deep-learning platform for medical imaging. *Biomedicine*, 158:113–122.

Giedd, J., Castellanos, F., Rajapakse, J., Vaituzis, A., and Rapoport, J. (1997). Sexual dimorphism of the developing human brain. *Prog Neuropsychopharmacol Biol Psychiatry*, 21:1185–1201.

Giedd, J., Raznahan, A., Mills, K., and Lenroot, R. (2012). Review: magnetic resonance imaging of male/female differences in human adolescent brain anatomy. *Biology of Sex Differences*, 3:19.

Glasser, M., Coalson, T., Robinson, E., Hacker, C., Harwell, J., Yacoub, E., Ugurbil, K. andersson, J., Beckmann, C., Jenkinson, M., Smith, S., and Van Essen, D. (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, 536:171–178.

Gobinath, A., Choleris, E., and Galea, L. (2017). Sex, hormones, and genotype interact to influence psychiatric disease, treatment, and behavioral research. *J Neurosci Res.*, 95:50–64.

Goddard, A., Mason, G., Almai, A., Rothman, D., Behar, K., Petroff, O., Charney, D., and Krystal, J. (2001). Reductions in occipital cortex GABA levels in panic disorder detected with H-1-magnetic resonance spectroscopy. *Archives of General Psychiatry*, 58:556–561.

Goelman, G., Dan, R., Røuˇ zička, F., Bezdicek, O., Røuˇ zička, E., Roth, J., Vymazal, J., and Jech, R. (2017). Frequency-phase analysis of resting-state functional MRI. *Nature Scientific Reports*, 7:43743.

Goldstein, J., Jerram, M., Abbs, B., Whitfield-Gabrieli, S., and Makris, N. (2010). Sex differences in stress response circuitry activation dependent on female hormonal cycle. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 30:431–438.

Goldstone, A., Mayhew, S., Przezdzik, I., Wilson, R., Hale, J., and Bagshaw, A. (2016). Gender Specific Re-organization of Resting-State Networks in Older Age. *Front Aging Neurosci*, 8:285.

Gong, G., He, Y., Chen, Z., and Evans, A. (2012). Convergence and divergence of thickness correlations with diffusion connections across the human cerebral cortex. *Neuroimage*, 59:1239–1248.

Gong, G., Rosa-Neto, P., Carbonell, F., Chen, Z., He, Y., and Evans, A. (2009). Age-and gender-related differences in the cortical anatomical network. *J Neurosci*, 29:15684–15693.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Nets. *Advances in Neural Information Processing Systems 27*, page 2672–2680.

Goodfellow, I., Shlens, J., and Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples. *ICLR 2015*.

Goto, M., Abe, O., Miyati, T., Yamasue, H., Gomi, T., and Takeda, T. (2016). Head Motion and Correction Methods in Resting-state Functional MRI. *Magn Reson Med Sci*, 15:178–186.

Gottesman, O., Johansson, F., Komorowski, M., Faisal, A., Sontag, D., Doshi-Velez, F., and Celi, L. (2019). Guidelines for reinforcement learning in healthcare. *Nature Medicine*, 25:16–18.

Goulden, N., Khusnulina, A., Davis, N., Bracewell, R., Bokde, A. L., McNulty, J., and Mullins, P. (2014). The salience network is responsible for switching between the default

mode network and the central executive network: Replication from DCM. *NeuroImage*, 99:180–190.

Goñi, J., Avena-Koenigsberger, A., de Mendizabal, N., Heuvel, M., Betzel, R., and Sporns, O. (2013a). Exploring the Morphospace of Communication Efficiency in Complex Networks. *PLoS One*, 8:e58070.

Goñi, J., Heuvel, M., Avena-Koenigsberger, A., Mendizabal, N., Betzel, R., Griffa, A., Hagmann, P., Corominas-Murtra, B., Thiran, J., and Sporns, O. (2013b). Resting-brain functional connectivity predicted by analytic measures of network communication. *PNAS*, 111:833–838.

Graham, J., Salimi-Khorshidi, G., Hagan, C., Walsh, N., Goodyer, I., Lennox, B., and Suckling, J. (2013). Meta-analytic evidence for neuroimaging models of depression: state or trait? *J Affect Disord.*, 151:423–431.

Grajauskas, L., Frizzell, T., Song, X., and D'Arcy, R. (2019). White Matter fMRI Activation Cannot Be Treated as a Nuisance Regressor: Overcoming a Historical Blind Spot. *Front. Neurosci.*, 13.

Greene, A., Gao, S., Scheinost, D., and Constable, R. (2018). Task-induced brain state manipulation improves prediction of individual traits. *Nature Communications*, 9.

Greenwood, E. (1945). *Experimental sociology: a study in method*. King's Crown Press, New York, 1st edition.

Greenwood, P. (2007). Functional plasticity in cognitive aging: review and hypothesis. *Neuropsychology*, 21:657–673.

Greicius, M., Krasnow, B., Reiss, A., and Menon, V. (2003). Functional connectivity in the resting brain: A network analysis of the default mode hypothesis. *PNAS*, 100:253–258.

Greicius, M., Supekar, K., Menon, V., and Dougherty, R. (2009). Resting-state functional connectivity reflects structural connectivity in the default mode network. *Cereb. Cortex*, 19:72–78.

Guimera, R., Sales-Pardo, M., and Amaral, L. (2007). Classes of complex networks defined by role-to-role connectivity profiles. *Nat Phys*, 3:63–69.

Guo, W., Liu, F., Liu, J., Yu, M., Zhang, Z., Liu, G., Xiao, C., and Zhao, J. (2015). Increased cerebellar-default-mode-network connectivity in drug-naive major depressive disorder at rest. *Medicine (Baltimore)*, 94:e560.

Gur, R., Alsop, D., Glahn, D., Petty, R., Swanson, C., Maldjian, J., Turetsky, B., Detre, J., Gee, J., and Gur, R. (2000). An fMRI study of sex differences in regional activation to a verbal and a spatial task. *Brain Lang*, 74:157–170.

Gur, R. and Gur, R. (2016). Sex differences in brain and behavior in adolescence: findings from the Philadelphia Neurodevelopmental Cohort. *Neurosci Biobehav Rev.*, 70:159–170.

Gusnard, D., Akbudak, E., Shulman, G., and Raichle, M. (2001). Medial prefrontal cortex and self-referential mental activity: Relation to a default mode of brain function. *PNAS*, 98:4259–4264.

Ha, S., Sohn, I., Kim, N., Sim, H., and Cheon, K. (2015). Characteristics of Brains in Autism Spectrum Disorder: Structure, Function and Connectivity across the Lifespan. *Exp Neurobiol*, 24:273–284.

Haar, S., Berman, S., Behrmann, M., and Dinstein, I. (2016). Anatomical Abnormalities in Autism? *Cerebral Cortex*, 26:1440–1452.

Hagan, C., Graham, J., Tait, R., Widmer, B., van Nieuwenhuizen, A., Ooi, C., Whitaker, K., Simas, T., Bullmore, E., Lennox, B., Sahakian, B., Goodyer, I., and Suckling, J. (2015). Adolescents with current major depressive disorder show dissimilar patterns of age-related differences in ACC and thalamus. *Neuroimage: Clinical*, 7:391–399.

Hagan, C., Graham, J., Widmer, B., Holt, R., Ooi, C., van Nieuwenhuizen, A., Fonagy, P., Reynolds, S., Target, M., Kelvin, R., Wilkinson, P., Bullmore, E., Lennox, B., Sahakian, B., Goodyer, I., and Suckling, J. (2013). Magnetic resonance imaging of a randomized controlled trial investigating predictors of recovery following psychological treatment in adolescents with moderate to severe unipolar depression: study protocol for Magnetic Resonance-Improving Mood with Psychoanalytic and Cognitive Therapies (MR-IMPACT). *BMC Psychiatry*, 13.

Hagmann, P., Cammoun, L., Gigandet, X., Meuli, R., Honey, C., Wedeen, V., and Sporns, O. (2008). Mapping the structural core of human cerebral cortex. *PLoS Biol*, 6:e159.

Hall, D., Huerta, M., McAuliffe, M., and Farber, G. (2012). Sharing Heterogeneous Data: The National Database for Autism Research. *Neuroinformatics*, 10:331–339.

Hamann, S., Herman, R., Nolan, C., and Wallen, K. (2004). Men and women differ in amygdala response to visual sexual stimuli. *Nat Neurosci*, 7:411–416.

Hamilton, W., Ying, R., and Leskovec, J. (2017). Representation Learning on Graphs: Methods and Applications. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*.

Han, S., Wuang, W., Zhang, Y., Zhao, J., and Chen, H. (2017). Recognition of early-onset schizophrenia using deep-learning method. *Applied Informatics*, 4.

Handwerker, D., Ollinger, J., and D'Esposito, M. (2004). Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses. *NeuroImage*, 21:1639–1651.

Happé, F. and Frith, U. (2006). The weak coherence account: detail-focused cognitive style in autism spectrum disorders. *J Autism Dev Disord*, 36:5–25.

Hariri, A., Tessitore, A., Mattay, V., Fera, F., and Weinberger, D. (2002). The amygdala response to emotional stimuli: a comparison of faces and scenes. *Neuroimage*, 17:317–323.

Hazlett, H., Gu, H., Munsell, B., Kim, S., Styner, M., Wolff, J., Elison, J., Swanson, M., Zhu, H., Botteron, K., Collins, D., Constantino, J., Dager, S., Estes, A., Evans, A., Fonov, V., Gerig, G., Kostopoulos, P., McKinstry, R., Pandey, J., Paterson, S., Pruett, J., Schultz, R., Shaw, D., Zwaigenbaum, L., and Piven, J. (2017). Early brain development in infants at high risk for autism spectrum disorder. *Nature*, 542:348–351.

He, B., Shulman, G., Snyder, A., and Corbetta, M. (2007). The role of impaired neuronal communication in neurological disorders. *Curr. Opin. Neurol.*, 20:655–660.

He, T., Kong, R., Holmes, A., Sabuncu, M., Eickhoff, M., Bzdok, D., Feng, J., and Yeo, B. (2018). Is deep learning better than kernel regression for functional connectivity prediction of fluid intelligence? *2018 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*.

He, Y., Wang, J., Wang, L., Chen, Z., Yan, C., Yang, H., Tang, H., Zhu, C., Gong, Q., Zang, Y., and Evans, A. (2009). Uncovering intrinsic modular organization of spontaneous brain activity in humans. *PLoS One*, 4:e5226.

Hechtlinger, Y., Chakravarti, P., and Qin, J. (2017). A Generalization of Convolutional Neural Networks to Graph-Structured Data. *arXiv*.

Heinsfeld, A., Franco, A., Craddock, R., Buchweitz, A., and Meneguzzia, F. (2018). Identification of autism spectrum disorder using deep learning and the ABIDE dataset. *NeuroImage: Clinical*, 17:16–23.

Hermundstad, A., Bassett, D., Brown, K., Aminoff, E., Clewett, D., Freeman, S., Frithsen, A., Johnson, A., Tipper, C., Miller, M., Grafton, S., and Carlson, J. (2013). Structural foundations of resting-state and task-based functional connectivity in the human brain. *PNAS*, 110:6169–6174.

Hilgetag, C., Burns, G., O'Neill, M., Scannell, J., and Young, M. (2000). Anatomical connectivity defines the organisation of cortical areas in the macaque monkey and the cat. *Phil. Trans. R. Soc. Lond.*, 355:91–110.

Hinton, G., Osindero, S., and Teh, Y.-H. (2006). A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, 18:1527–1554.

Ho, D., Imai, K., King, G., and Stuart, E. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15:199–236.

Hofer, A., Siedentopf, C., Ischebeck, A., et al. (2006). Gender differences in regional cerebral activity during the perception of emotion: a functional MRI study. *Neuroimage*, 32:854–862.

Hollander, E., Anagnostou, E., Chaplin, W., Esposito, K., Haznedar, M., Licalzi, E., Wasserman, S., Soorya, L., and Buchsbaum, M. (2005). Striatal volume on magnetic resonance imaging and repetitive behaviors in autism. *Biol Psychiatry*, 58:226–232.

Honey, C., Sporns, O., Cammoun, L., Gigandet, X., Thiran, J., Meuli, R., and Hagmann, P. (2009). Predicting human resting-state functional connectivity from structural connectivity. *PNAS*, 106:2035–2040.

Horn, A., Ostwald, D., Reisert, M., and Blankenburg, F. (2013). The structural-functional connectome and the default mode network of the human brain. *NeuroImage*, 102 (Pt 1):142–151.

Hromkovic, J., Klasing, R., Pelc, A., Ruzicka, P., and Unger, W. (2005). *Dissemination of Information in Communication Networks: Broadcasting, Gossiping, Leader Election, and Fault-Tolerance*. Springer-Verlag, Berlin, Heidelberg.

Hua, C., Wang, H., Wang, H., Lu, S., Liu, C., and Khalid, S. (2019). A Novel Method of Building Functional Brain Network using Deep Learning Algorithm with Application in Proficiency Detection. *International Journal of Neural Systems*, 29:1850015.

Hugdahl, K., Thomsen, T., and Ersland, L. (2006). Sex differences in visuo-spatial processing: an fMRI study of mental rotation. *Neuropsychologia*, 44:1575–1583.

Hull, J., Dokovna, L., Jacokes, Z., Torgerson, C., Irimia, A., and van Horn, J. (2017). Resting-State Functional Connectivity in Autism Spectrum Disorders: A Review. *Front Psychiatry*, 7.

Hurlburt, R., Alderson-Day, B., K uhn, S., and Fernyhough, C. (2016). Exploring the Ecological Validity of Thinking on Demand: Neural Correlates of Elicited vs. Spontaneously Occurring Inner Speech. *PLoS ONE*, 11:e0147932.

Hussain, Z., Gimenez, F., Yi, D., and Rubin, D. (2017). Differential Data Augmentation Techniques for Medical Imaging Classification Tasks. *AMIA Annu Symp Proc*, 2017:979–984.

Hutchison, R., Womelsdorf, T., Allen, E., Bandettini, P., Calhoun, V., Corbetta, M., Della Penna, S., Duyn, J., Glover, G., Gonzalez-Castillo, J., Handwerker, D., Keilholz, S., Kiviniemi, V., Leopold, D., de Pasquale, F., Sporns, O., Walter, M., and Chang, C. (2013). Dynamic functional connectivity: promise, issues, and interpretations. *NeuroImage*, 80:360–378.

H utt, M.-T., Kaiser, M., and Hilgetag, C. (2014). Perspective: Network-guided pattern formation of neural dynamics. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 369:20130522.

Iidaka, T. (2015). Resting state functional magnetic resonance imaging and neural network classified autism and control. *Cortex*, 63:55–67.

Imbens, G. (2004). Nonparametric estimation of average treatment effects under exogeneity: a review. *Review of Economics and Statistics*, 86:4–29.

Ioffe, S. and Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv*.

Ito, T., Kulkarni, K., Schultz, D., Mill, R., Chen, R., Solomyak, L., and Cole, M. (2017). Cognitive task information is transferred between brain regions via resting-state network topology. *Nature Communications*, 8:1027.

Jang, H., Plis, S., Calhoun, V., and Lee, J. (2017). Task-specific feature extraction and classification of fMRI volumes using a deep neural network initialized with a deep belief network: Evaluation using sensorimotor tasks. *Neuroimage*, 145(Pt B):314–328.

Ji, G., Zhang, Z., Xu, Q., Zang, Y., Liao, W., and Lu, G. (2014). Generalized tonic-clonic seizures: aberrant interhemispheric functional and anatomical connectivity. *Radiology*, 271:839–847.

Jie, B., Zhang, D., Wee, C.-Y., and Shen, D. (2013). Topological graph kernel on multiple thresholded functional connectivity networks for mild cognitive impairment classification. *Hum Brain Mapp*, 35:2876–2897.

Johnston, J., Vaishnavi, S., Smyth, M., Zhang, D., He, B., Zempel, J., Shimony, J., Snyder, A., and Raichle, M. (2008). Loss of resting interhemispheric functional connectivity after complete section of the corpus callosum. *J Neurosci.*, 28:6453–6458.

Jones, D., Kn osche, T., and Turner, R. (2013). White matter integrity, fiber count, and other fallacies: The do's and don'ts of diffusion MRI. *NeuroImage*, 73:239–254.

Jones, T., Bandettini, P., Kenworthy, L., Case, L., Milleville, S., Martin, A., and Birn, R. (2010). Sources of group differences in functional connectivity: an investigation applied to autism spectrum disorder. *NeuroImage*, 49:401–414.

Jonsson, B., Bjornsdottir, G., Thorgeirsson, T., Ellingsen, L., Walters, G., Gudbjartsson, D., Stefansson, H., Stefansson, K., and Ulfarsson, M. (2019). Brain age prediction using deep learning uncovers associated sequence variants. *Nat Commun*, 10.

Joyce, K., Laurienti, P., Burdette, J., and Hayasaka, S. (2010). A new measure of centrality for brain networks. *PLoS ONE*, 5:e12200.

Jung, M., Kosaka, H., Saito, D., Ishitobi, M., Morita, T., Inohara, K., Asano, M., Arai, S., Munesue, T., Tomoda, A., Wada, Y., Sadato, N., Okazawa, H., and Iidaka, T. (2014). Default mode network in young male adults with autism spectrum disorder: relationship with autism spectrum traits. *Mol Autism*, 5:35.

Just, M., Cherkassky, V., Keller, T., and Minshew, N. (2004). Cortical activation and synchronization during sentence comprehension in high-functioning autism: evidence of underconnectivity. *Brain*, 127:1811–21.

Kaiser, R.H. andrews-Hanna, J., Wager, T., and Pizzagalli, D. (2015). Large-Scale Network Dysfunction in Major Depressive Disorder: A Meta-analysis of Resting-State Functional Connectivity. *JAMA Psychiatry*, 72:603–611.

Kaiser, M. and Hilgetag, C. (2004). Edge vulnerability in neural and metabolic networks. *Biol. Cybern.*, 90:311–317.

Kaiser, M., Martin, R. andras, P., and Young, M. (2007). Simulation of robustness against lesions of cortical networks. *Eur J Neurosci*, 25:3185–3192.

Kallus, N. (2018). DeepMatch: Balancing Deep Covariate Representations for Causal Inference Using Adversarial Training. *arXiv*.

Karpathy, A. and Fei-Fei, L. (2014). Deep Visual-Semantic Alignments for Generating Image Descriptions. *CVPR 2015*.

Katuwal, G., Cahill, N., Baum, S., and Michael, A. (2015). The predictive power of structural MRI in Autism diagnosis. *Conf Proc IEEE Eng Med Biol Soc*, page 4270–4273.

Kawahara, J., Brown, C., Miller, S., Booth, B., Chau, V., Grunau, R., Zwicker, J., and Hamarneh, G. (2017). BrainNetCNN: Convolutional neural networks for brain networks; towards predicting neurodevelopment. *NeuroImage*, 146:1038–1049.

Kazeminejad, A. and Sotero, R. (2019). Topological Properties of Resting-State fMRI Functional Networks Improve Machine Learning-Based Autism Classification. *Front. Neurosci.*, 12.

Kennedy, D. and Courchesne, E. (2008). The intrinsic functional organization of the brain is altered in autism. *NeuroImage*, 39:1877–85.

Khosla, M. an Jamison, K., Kuceyeski, A., and Sabuncu, M. (2018). 3D Convolutional Neural Networks for Classification of Functional Connectomes. *MICCAI 2018*.

Khundrakpam, B., Lewis, J., Kostopoulos, P., Carbonell, F., and Evans, A. (2017). Cortical Thickness Abnormalities in Autism Spectrum Disorders Through Late Childhood, Adolescence, and Adulthood: A Large-Scale MRI Study. *Cerebral Cortex*, 27:1721–1731.

Kipf, T. and Welling, M. (2017). Semi-Supervised Classification with Graph Convolutional Neural Networks. *ICLR 2017*.

Klein, S., Smolka, M., Wrase, J., et al. (2003). The influence of gender and emotional valence of visual cues on FMRI activation in humans. *Pharmacopsychiatry*, 36:S191–S194.

Koch, M., Norris, D., and Hund-Georgiadis, M. (2002). An investigation of functional and anatomical connectivity using magnetic resonance imaging. *NeuroImage*, 16:241–250.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Intelligence - Volume 2, IJCAI'95, Morgan Kaufmann Publishers Inc., San Francisco, C*, pages 1137–1143.

Kong, X., Liu, Z., Huang, L., Wang, X., Yang, Z., Zhou, G., Zhen, Z., and Liu, J. (2015). Mapping individual brain networks using statistical similarity in regional morphology from MRI. *PLoS One*, 10:e0141840.

Kong, X., Wang, X., Huang, L., Pu, Y., Yang, Z., Dang, X., Zhen, Z., and Liu, J. (2014). Measuring individual morphological relationship of cortical regions. *J. Neurosci. Methods*, 237:103–107.

Kotikalapudi, R. and contributors (2017). keras-vis. `https://github.com/raghakot/keras-vis`.

Koyamada, S., Shikauchi, Y., Nakae, K., Koyama, M., and Ishii, S. (2015). Deep learning of fMRI big data: a novel approach to subject-transfer decoding. *arXiv*.

Kriege, N., Johansson, F., and Morris, C. (2019). A Survey on Graph Kernels. *arXiv*.

Kriston, K. (2011). Functional and Effective Connectivity: A Review. *Brain Connectivity*, 1:13–36.

Krizhevsky, A., Sutskever, I., and Hinton, G. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*.

Kundu, P., Brenowitz, N., Voon, V., Worbe, Y., Vèrtes, P., Inati, S., Saad, Z., Bandettini, P., and Bullmore, E. (2013). Integrated strategy for improving functional connectivity mapping using multiecho fMRI. *PNAS*, 110:16187–16192.

Kundu, P., Inati, S., Evans, J., Luh, W., and Bandettini, P. (2012). Differentiating BOLD and non-BOLD signals in fMRI time series using multi-echo EPI. *NeuroImage*, 60:1759–1770.

Kvalseth, T. (2017). On Normalized Mutual Information: Measure Derivations and Properties. *Entropy*, 19:631.

Lai, M., Lombardo, M., and Baron-Cohen, S. (2014). Autism. *Lancet*, 383:896–910.

Lange, N., Travers, B., Bigler, E., Prigge, M., Froehlich, A., Nielsen, J., et al. (2015). Longitudinal volumetric brain changes in autism spectrum disorder ages 6-35 years. *Autism Res*, 8:82–93.

Latora, V. and Marchiori, M. (2001). Efficient Behavior of Small-World Networks. *Phys. Rev. Lett.*, 87:198701.

Lauterbur, P. (1973). Image formation by induced local interactions: examples employing nuclear magnetic resonance. *Nature*, 242:190–191.

LeCun, Y., Haffner, P., Bottou, L., and Bengio, Y. (1999). Object Recognition with Gradient-Based Learning. *Lecture Notes in Computer Science*, 1681:319–345.

Leech, R. and Sharp, D. (2014). The role of the posterior cingulate cortex in cognition and disease. *Brain*, 137:12–32.

Leming, M. and Suckling, J. (2019). Deep Learning on Brain Images in Autism: What Do Large Samples Reveal of Its Complexity? In Ferrandez Vicente, J., Alvarez-Sanchez, J., de la Paz Lopez, F., Toledo Moreo, J., , and Adeli, H., editors, *Understanding the Brain Function and Emotions,* Proceedings of the 8th International Work-Conference on the Interplay Between Natural and Artificial Computation, Part I, pages 389–402. Springer.

Leming, M. and Suckling, J. (2020a). Ensemble deep learning on large, mixed-site fMRI datasets in autism and other tasks. *International Journal of Neural Systems*, 30:2050012–1–16.

Leming, M. and Suckling, J. (2020b). Stochastic encoding of graphs in deep learning allows for complex analysis of gender classification in resting-state and task functional brain networks from the UK Biobank. *IEEE (submitted).*

Lenroot, R. and Giedd, J. (2006). Brain development in children and adolescents: insights from anatomical magnetic resonance imaging. *Neurosci. Biobehav. Rev.*, 30:718–729.

Leulescu, A. and Agafitei, M. (2013). Statistical matching: a model based approach for data integration. Technical report, European Commission.

Levy, F. (2007). Theories of autism. *Aust N Z J Psychiatry*, 41:859–868.

Li, B., Liu, L., Friston, K., Shen, H., Wang, L., Zeng, L., and Hu, D. (2013). A Treatment-Resistant Default Mode Subnetwork in Major Depression. *Biological Psychiatry*, 74:48–54.

Li, G. and Yu, Y. (2018). Contrast-Oriented Deep Neural Networks for Salient Object Detection. *IEEE Transactions on Neural Networks and Learning Systems*, 29:6038–6051.

Li, H. and Fan, Y. (2018). Brain decoding from functional MRI using long short-term memory recurrent neural networks. *arXiv.*

Li, K., Guo, L., Nie, J., Li, G., and Liu, T. (2009). Review of methods for functional brain connectivity detection using fMRI. *Computerized medical imaging and graphics: the official journal of the Computerized Medical Imaging Society*, 33:131–139.

Lipton, Z. (2015). A Critical Review of Recurrent Neural Networks for Sequence Learning. *arXiv*.

Litjens, G., Kooi, T., Bejnordi, B., Setio, A., Ciompi, F., Ghafoorian, M., van der Laak, J., van Ginneken, B., and Sánchez, C. (2017). A Survey on Deep Learning in Medical Image Analysis. *Medical Image Analysis*, 42:60–88.

Liu, B., Wei, Y., Zhang, Y., and Yang, Q. (2014). Deep Neural Networks for High Dimension, Low Sample Size Data. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 2287–2293.

Liu, H., Stufflebeam, S., Sepulcre, J., Hedden, T., and Buckner, R. (2009). Evidence from intrinsic activity that asymmetry of the human brain is controlled by multiple factors. *PNAS*, 106:20499–20503.

Liu, Y., Liang, M., Zhou, Y., He, Y., Hao, Y., Song, M., Yu, C., Liu, H., Liu, Z., and Jiang, T. (2008). Disrupted small-world networks in schizophrenia. *Brain*, 131:945–961.

Lo, Y.-C., Soong, W.-T., Gau, S.-F., Wu, Y.-Y., Lai, M.-C., Yeh, F.-C., Chiang, W.-Y., Kuo, L.-W., Jaw, F.-S., and Tseng, W.-Y. (2011). The loss of asymmetry and reduced interhemispheric connectivity in adolescents with autism: A study using diffusion spectrum imaging tractography. *Psychiatry Res.*, 192:60–66.

Lohmann, G., Margulies, D., Horstmann, A., Pleger, B., Lepsien, J., Goldhahn, D., H., S., Stumvoll, M., Villringer, A., and Turner, R. (2010). Eigenvector centrality mapping for analyzing connectivity patterns in fMRI data of the human brain. *PLoS ONE*, 5:e10232.

Lopez-Larson, M.P. anderson, J., Ferguson, M., and Yurgelun-Todd, D. (2011). Local brain connectivity and associations with gender and age. *Dev Cogn Neurosci*, 1:187–197.

Lord, C., Cook, E., Leventhal, B., and Amaral, D. (2000). Autism spectrum disorders. *Neuron*, 28:355–363.

Lynall, M., Bassett, D., Kerwin, R., McKenna, P., Kitzbichler, M., Muller, U., and Bullmore, E. (2010). Functional connectivity and brain networks in schizophrenia. *J Neurosci*, 30:9477–9487.

Mackiewicz, K., Sarinopoulos, I., Cleven, K., and Nitschke, J. (2006). The effect of anticipation and the specificity of sex differences for amygdala and hippocampus function in emotional memory. *PNAS*, 103:14200–14205.

Mansfield, P. and Maudsley, A. (1977). Medical imaging by NMR. *Br J Radiol*, 50:188–194.

Masuda, N. and Aihara, K. (2004). Global and local synchrony of coupled neurons in small-world networks. *Biol. Cybern*, 90:302–309.

Maturana, D. and Scherer, S. (2015). VoxNet: A 3D Convolutional Neural Network for Real-Time Object Recognition. *Intelligent Robots and Systems (IROS)*, pages 922–928.

Mazoyer, B., Zago, L., Mellet, E., Bricogne, S., Etard, O., Houdè, O., Crivello, F., Joliot, M., Petit, L., and Tzourio-Mazoyer, N. (2001). Cortical networks for working memory and executive functions sustain the conscious resting state in man. *Brain Res Bull.*, 54:287–298.

Mazure, C. and Swendsen, J. (2016). Sex differences in Alzheimer's disease and other dementias. *Lancet Neurol.*, 15:451–452.

McAlonan, G., Cheung, V., Cheung, C., Suckling, J., Lam, G., Tai, K., Yip, L., Murphy, D., and Chua, S. (2005). Mapping the brain in autism. A voxel-based MRI study of volumetric differences and intercorrelations in autism. *Brain*, 128(Pt 2):268–276.

McAlonan, G., Daly, E., Kumari, V., Critchley, H., van Amelsvoort, T., Suckling, J., Simmons, A., Sigmundsson, T., Greenwood, K., Russell, A., Schmitz, N., Happe, F., Howlin, P., and Murphy, D. (2002). Brain anatomy and sensorimotor gating in Asperger's syndrome. *Brain*, 125:1594–1606.

McClure, E., Monk, C., Nelson, E., et al. (2004). A developmental examination of gender differences in brain engagement during evaluation of threat. *Biological Psychiatry*, 55:1047–1055.

McKiernan, K., Kaufman, J., Kucera-Thompson, J., and Binder, J. (2003). A parametric manipulation of factors affecting task-induced deactivation in functional neuroimaging. *J. Cognit. Neurosci.*, 15:394–408.

Mechelli, A., Friston, K., Frackowiak, R., and Price, C. (2005). Structural covariance in the human cortex. *J. Neurosci.*, 25:8303–8310.

Meier, J., Topka, M., and H anggi, J. (2016). Differences in Cortical Representation and Structural Connectivity of Hands and Feet between Professional Handball Players and Ballet Dancers. *Neural Plast.*, 2016.

Meszlényi, R., Buza, K., and Vidnyánszky, Z. (2017). Resting state fMRI functional connectivity-based classification using a convolutional neural network architecture. *Front Neuroinform.*, 11:61.

Milgram, S. (1967). The Small-World Problem. *Psychology Today*, 1:61–67.

Mitra, A. and Raichle, M. (2016). How networks communicate: propagation patterns in spontaneous brain activity. *Philos Trans R Soc Lond B Biol Sci.*, 371:20150546.

Mi˘ sić, B., Betzel, R., Nematzadeh, A., Goñi, J., Griffa, A., Hagmann, P., and Sporns, O. (2015). Cooperative and competitive spreading dynamics on the human connectome. *Neuron*, 86:1518–1529.

Modinos, G., Mechelli, A., Pettersson-Yeo, W., Allen, P., McGuire, P., and Aleman, A. (2013). Pattern classification of brain activation during emotional processing in subclinical depression: psychosis proneness as potential confounding factor. *Peerj.*, 1:e42.

Morgan, S. and Harding, D. (2006). Matching estimators of causal effects: Prospects and pitfalls in theory and practice. *Sociological Methods and Research*, 35:3–60.

Mucha, P., Richardson, T., Macon, K., Mason, A. Porter, M., and Onnela, J. (2010). Community Structure in Time-Dependent, Multiscale, and Multiplex Networks. *Science*, 328:876–878.

Mueller, S., Wang, D., Fox, M., Yeo, B., Sepulcre, J., Sabuncu, M., Shafee, R., Lug, J., and Liu, H. (2013). Individual Variability in Functional Connectivity Architecture of the Human Brain. *Neuron*, 77:586–595.

Mulders, P., van Eijndhoven, P., Schene, A., Beckmann, C., and Tendolkar, I. (2015). Resting-state functional connectivity in major depressive disorder: A review. *Neurosci Biobehav Rev.*, 56:330–344.

Müller, E., Schuler, A., and Yates, G. (2008). Social challenges and supports from the perspective of individuals with Asperger syndrome and other autism spectrum disabilities. *Autism*, 12:173–190.

Munir, K., Elahi, H., Ayub, A., Frezza, F., and Rizzi, A. (2019). Cancer Diagnosis Using Deep Learning: A Bibliographic Review. *Cancers*, 11:1235.

Murphy, K., Birn, R., Handwerker, D., Jones, T., and Bandettini, P. (2009). The impact of global signal regression on resting state correlations: are anti-correlated networks introduced? *NeuroImage*, 44:893–905.

Nakamura, T., Hillary, F., and Biswal, B. (2009). Resting network plasticity following brain injury. *PLoS One*, 4:e8220.

Nebel, M., Eloyan, A., Barber, A., and Mostofsky, S. (2014a). Precentral gyrus functional connectivity signatures of autism. *Front Syst Neurosci*, 8:80.

Nebel, M., Joel, S., Muschelli, J., Barber, A., Caffo, B., Pekar, J., and Mostofsky, S. (2014b). Disruption of functional organization within the primary motor cortex in children with autism. *Hum Brain Mapp*, 35:567–580.

Nickl-Jockschat, T., Habel, U., Michel, T., Manning, J., Laird, A., Fox, P., Schneider, F., and Eickhoff, S. (2012a). Brain structure anomalies in autism spectrum disorder–a meta-analysis of VBM studies using anatomic likelihood estimation. *Hum Brain Mapp*, 33:1470–1489.

Nickl-Jockschat, T., Habel, U., Michel, T., Manning, J., Laird, A., Fox, P., Schneider, F., and Eickhoff, S. (2012b). Brain Structure Anomalies in Autism Spectrum Disorder—A Meta-Analysis of VBM Studies Using Anatomic Likelihood Estimation. *Hum Brain Mapp*, 33:1470–1489.

Nicolini, C., Bordier, C., and Bifone, A. (2017). Community detection in weighted brain connectivity networks beyond the resolution limit. *NeuroImage*, 146:28–39.

Nielsen, J., Zielinski, B., Ferguson, M., and Lainhart, J.E. anderson, J. (2013a). An evaluation of the left-brain vs. right-brain hypothesis with resting state functional connectivity magnetic resonance imaging. *PLoS One*, 8:e71275.

Nielsen, J., Zielinski, B., Fletcher, P., Alexander, A., Lange, N., Bigler, E., and Lainhart, J.E. anderson, J. (2013b). Multisite functional connectivity MRI classification of autism: ABIDE results. *Front Hum Neurosci*, 7:599.

Nikolentzos, G., Meladianos, P., Tixier, A., Skianis, K., and Vazirgiannis, M. (2017). Kernel Graph Convolutional Neural Networks. *ICANN 2018*.

Nikolentzos, G., Siglidis, G., and Vazirgiannis, M. (2019). Graph Kernels: A Survey. *arXiv*.

Nord, C., Valton, V., Wood, J., and Roiser, J. (2017). Power-up: A Reanalysis of 'Power Failure' in Neuroscience Using Mixture Modeling. *Journal of Neuroscience*, 37:8051–8061.

O'Dwyer, L., Tanner, C., van Dongen, E., Greven, C., Bralten, J., Zwiers, M., Franke, B., Heslenfeld, D., Oosterlaan, J., Hoekstra, P., Hartman, C., Groen, W., Rommelse, N.,

and Buitelaar, J. (2016). Decreased Left Caudate Volume Is Associated with Increased Severity of Autistic-Like Symptoms in a Cohort of ADHD Patients and Their Unaffected Siblings. *PLoS One*, 11:e0165620.

Ogawa, S., Lee, T., Kay, A., and Tank, D. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *PNAS*, 87:9868–9872.

Opitz, D. and Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, 11:169–198.

Oquab, M., Bottou, L., Laptev, I., and Sivic, J. (2014). Learning and transferring mid-level image representations using convolutional neural networks. *Proc. CVPR*.

Oquab, M., Bottou, L., Laptev, I., and Sivic, J. (2015). Is object localization for free? weakly-supervised learning with convolutional neural networks. *Proc. CVPR*.

Papo, D., Buldù, J., Boccaletti, S., and Bullmore, E. (2014). Complex Network Theory and the Brain. *Philos Trans R Soc Lond B Biol Sci*, 369:20130520.

Paquola, C., Vos De Wael, R., Wagstyl, K., Bethlehem, R., Hong, S.-J., Seidlitz, J., Bullmore, E., Evans, A., Misic, B., Margulies, D., Smallwood, J., and Bernhardt, B. (2019). Microstructural and functional gradients are increasingly dissociated in transmodal cortices. *PLoS Biology*, 17:e3000284.

Park, H. and Friston, K. (2013). Structural and functional brain networks: from connections to cognition. *Science*, 342:1238411.

Passingham, R., Stephan, K., and K otter, R. (2002). The anatomical basis of functional localization in the cortex. *Nat. Rev. Neurosci.*, 3:606–616.

Patel, A. and Bullmore, E. (2016). A wavelet-based estimator of the degrees of freedom in denoised fMRI time series for probabilistic testing of functional connectivity and brain graphs. *NeuroImage*, 142:14–26.

Patel, A., Kundu, P., Rubinov, M., Jones, P., Vertes, P., Ersche, K., Suckling, J., and Bullmore, E. (2014). A wavelet method for modeling and despiking motion artifacts from resting-state fMRI time series. *NeuroImage*, 95:287–304.

Pauling, L. and Coryell, C. (1936). The Magnetic Properties and Structure of Hemoglobin, Oxyhemoglobin and Carbonmonoxyhemoglobin. *PNAS*, 22:210–216.

Perone, C. and Cohen-Adad, J. (2019). Promises and limitations of deep learning for medical image segmentation. *Journal of Medical Artificial Intelligence*, 2.

Pinheiro, P. and Collobert, R. (2015). From image-level to pixel-level labeling with convolutional networks. *CVPR*.

Pizoli, C., Shah, M., Snyder, A., Shimony, J., Limbrick, D., Raichle, M., Schlaggar, B., and Smyth, M. (2011). Resting-state activity in development and maintenance of normal brain function. *PNAS*, 108:11638–11643.

Plitt, M., Barnes, K., and Martin, A. (2015). Functional connectivity classification of autism identifies highly predictive brain features but falls short of biomarker standards. *NeuroImage Clinical*, 7:359–66.

Poldrack, R., Barch, D., Mitchell, J., Wager, T., Wagner, A., Devlin, J., Cumba, C., Koyejo, O., and Milham, M. (2013). Toward open sharing of task-based fMRI data: the OpenfMRI project. *Front Neuroinform*, 7.

Poldrack, R. and Gorgolewski, K. (2017). OpenfMRI: Open sharing of task fMRI data. *NeuroImage*, 144:259–261.

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, 6:21–45.

Power, J., Cohen, A., Nelson, S., Wig, G., Barnes, K., Church, J., Vogel, A., Laumann, T., Miezin, F., Schlaggar, B., and Petersen, S. (2011). Functional network organization of the human brain. *Neuron*, 72:665–678.

Preston, D. (2006). Magnetic Resonance Imaging (MRI) of the Brain and Spine: Basics.

Preti, M., Bolton, T., and Van De Ville, D. (2017). The dynamic functional connectome: State-of-the-art and perspectives. *NeuroImage*, 160:41–54.

Price, T., Wee, C., Gao, W., and Shen, D. (2014). Multiple-network classification of childhood autism using functional connectivity dynamics. *Med Image Comput Comput Assist Interv*, 17:177–184.

Prigge, M., Bigler, E., Fletcher, P., Zielinski, B., Ravichandran, C., Anderson, J., Froehlich, A., Abildskov, T., Papadopolous, E., Maasberg, K., Nielsen, J., Alexander, A., Lange, N., and Lainhart, J. (2013). Longitudinal Heschl's gyrus growth during childhood and adolescence in typical development and autism. *Autism Res*, 6:78–90.

Purcell, E., Torrey, H., and Pound, R. (1946). Resonance Absorption by Nuclear Magnetic Moments in a Solid. *Phys. Rev.*, 69.

Putnam, M., Wig, G., Grafton, S., Kelley, W., and Gazzaniga, M. (2008). Structural organization of the corpus callosum predicts the extent and impact of cortical activity in the nondominant hemisphere. *J Neurosci.*, 28:2912–2918.

Qiu, A., Anh, T., Li, Y., Chen, H., Rifkin-Graboi, A., Broekman, B., Kwek, K., Saw, S., Chong, Y., Gluckman, P., Fortier, M., and Meaney, M. (2015). Prenatal maternal depression alters amygdala functional connectivity in 6-month-old infants. *Transl Psychiatry*, 5:e508.

Qiu, T., Chang, C., Li, Y., Qian, L., Xiao, C., Xiao, T., Xiao, X., Xiao, Y., Chu, K., Lewis, M., and Ke, X. (2016). Two years changes in the development of caudate nucleus are involved in restricted repetitive behaviors in 2–5-year-old children with autism spectrum disorder. *Developmental Cognitive Neuroscience*, 19:137–143.

Quach, K. (2018). IBM Watson dishes out 'dodgy cancer advice', Google Translate isn't better than humans yet, and other AI tidbits. *The Register*.

Quigley, M., Cordes, D., Turski, P., Moritz, C., Haughton, V., Seth, R., and Meyerand, M. (2003). Role of the corpus callosum in functional connectivity. *AJNR Am J Neuroradiol*, 24:208–212.

Raichle, M., MacLeod, A., Snyder, A., Powers, W., Gusnard, D., and Shulman, G. (2001). A default mode of brain function. *PNAS*, 98:676–682.

Ramasubbu, R., Brown, M., Cortese, F., Gaxiola, I., Goodyear, B., Greenshaw, A., Dursun, S., and Greiner, R. (2016). Accuracy of automated classification of major depressive disorder as a function of symptom severity. *NeuroImage: Clinical*, 12:320–331.

Ramdas, A., Garcia, N., and Cuturi, M. (2017). On Wasserstein Two Sample Testing and Related Families of Nonparametric Tests. *Entropy*, 19:47.

Redcay, E. and Courchesne, E. (2005). 2005. *Biol Psychiatry.*, 58:1–9.

Ribeiro, M., Singh, S., and Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. *Knowledge Discovery and Data Mining (KDD)*.

Ritchie, S., Cox, S., Shen, X., Lombardo, M., Reus, L., Alloza, C., Harris, M., Alderson, H., Hunter, S., Neilson, E., Liewald, D., Auyeung, B., Whalley, H., Lawrie, S., Gale, C.,

Bastin, M., McIntosh, A., and Deary, I. (2018). Sex Differences in the Adult Human Brain: Evidence from 5216 UK Biobank Participants. *Cerebral Cortex*, 28:2959–2975.

Rocha, L. (2002). Proximity and Semi-Metric Analysis of Social Networks: Advanced Knowledge Integration In Assessing Terrorist Threats. Technical Report LAUR 02–6557, Los Aalamos National Laboratory, Los Alamos, New Mexico.

Rogers, S., Vismara, L., Wagner, A., McCormick, C., Young, G., and Ozonoff, S. (2014). Autism Treatment in the First Year of Life: A Pilot Study of Infant Start, a Parent-Implemented Intervention for Symptomatic Infants. *Journal of Autism and Developmental Disorders*, 44:2981–2995.

Roiser, J. and Sahakian, B. (2013). Hot and cold cognition in depression. *CNS Spectr.*, 18:139–149.

Rojas, D., Peterson, E., Winterrowd, E., Reite, M., Rogers, S., and Tregellas, J. (2006). Regional gray matter volumetric changes in autism associated with social and repetitive behavior symptoms. *BMC Psychiatry*, 6:56.

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, 33:1–39.

Romero-Munguía, M. (2013). Theory of mind deficit versus faulty procedural memory in autism spectrum disorders. *Autism Res Treat*, 2013:128264.

Rosa, M., Portugal, L., Hahn, T., Fallgatter, A., Garrido, M., Shawe-Taylor, J., and Mourao-Miranda, J. (2015). Sparse network-based models for patient classification using fMRI. *NeuroImage*, 105:493–506.

Rosenbaum, P. (1989). Optimal Matching for Observational Studies. *Journal of the American Statistical Association*, 84:1024–1042.

Rubin, D. (1973). Matching to remove bias in observational studies. *Biometrics*, 29:159–184.

Rubinov, M. and Sporns, O. (2010). Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage*, 52:1059–1069.

Rubner, Y., Tomasi, C., and Guibas, L. (2000). The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision*, 40:99–121.

Rudie, J., Brown, J., Beck-Pancer, D., Hernandez, L., Dennis, E., Thompson, P., Bookheimer, S., and Dapretto, M. (2013). Altered functional and structural brain network organization in autism. *NeuroImage: Clinical*, 2:79–94.

Ruigrok, A., Salimi-Khorshidi, G., Lai, M., Baron-Cohen, S., Lombardo, M., Tait, R., and Suckling, J. (2014). A meta-analysis of sex differences in human brain structure. *Neuroscience and Biobehavioral Reviews*, 39:34–50.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A., and Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115:211–252.

Rutter, M., Caspi, A., and Moffitt, T. (2003). Using sex differences in psy- chopathology to study causal mechanisms: unifying issues and research strategies. *J Child Psychol Psychiatry*, 44:1092–1115.

Sacher, J., Neumann, J., Okon-Singer, H., Gotowiec, S., and Villringer, A. (2013a). Sexual dimorphism in the human brain: evidence from neuroimaging. *Magn. Reson. Imaging*, 31:366–375.

Sacher, J., Okon-Singer, H., and Villringer, A. (2013b). Evidence from neuroimaging for the role of the menstrual cycle in the interplay of emotion and cognition. *Front Hum Neurosci*, 7.

Salvador, R., Martinez, A., Pomarol-Clotet, E., Gomar, J., Vila, F., Sarro, S., Capdevila, A., and Bullmore, E. (2008). A simple view of the brain through a frequency-specific functional connectivity measure. *Neuroimage*, 39:279–289.

Salvador, R., Suckling, J., Schwarzbauer, C., and Bullmore, E. (2005). Undirected graphs of frequency-dependent functional connectivity in whole brain networks. *Philos Trans R Soc Lond B Biol Sci*, 360:937–946.

Sato, J., Moll, J., Green, S., Deakin, J., Thomaz, C., and Zahn, R. (2015). Machine learning algorithm accurately detects fMRI signature of vulnerability to major depression. *Psychiatry Research: Neuroimaging*, 233:289–291.

Satterthwaite, T., Wolf, D., Roalf, D., Ruparel, K., Erus, G., Vandekar, S., Gennatas, E., Elliott, M., Smith, A., Hakonarson, H., Verma, R., Davatzikos, C., Gur, R., and Gur, R. (2015). Linked Sex Differences in Cognition and Functional Connectivity in Youth. *Cereb Cortex*, 25:2383–2394.

Schaper, F.L.W.V.J. Zhao, Y., Janssen, M., Wagner, G., Colon, A., Hilkman, D., Gommer, E., Vlooswijk, M., Hoogland, G., Ackermans, L., Bour, L., Van Wezel, R., Boon, P., Temel, Y., Heida, T., van Kranen-Mastenbroek, V., and Rouhl, R. (2019). Single cell recordings

to target the anterior nucleus of the thalamus in deep brain stimulation for patients with refractory epilepsy. *International Journal of Neural Systems*, 29.

Schienle, A., Sch afer, A., Stark, R., Walter, B., and Vaitl, D. (2005). Gender differences in the processing of disgust- and fear-inducing pictures: an fMRI study. *Neuroreport*, 16:277–280.

Schmidt, R., LaFleur, K., de Reus, M., van den Berg, L., and van den Heuvel, M. (2015). Kuramoto model simulation of neural hubs and dynamic synchrony in the human cerebral connectome. *BMC Neuroscience*, 16:54.

Schmitt, D., Realo, A., Voracek, M., and Allik, J. (2008). Why can't a man be more like a woman? Sex differences in Big Five personality traits across 55 cultures. *J Pers Soc Psychol.*, 94:168–182.

Schulz, M., Yeo, T., Vogelstein, J., Mourao-Miranada, J., Kather, J., Kording, K., Richards, B., and Bzdok, D. (2019). Deep learning for brains?: Different linear and nonlinear scaling in UK Biobank brain images vs. machine-learning datasets. *bioRxiv*.

Schumann, C., Hamstra, J., Goodlin-Jones, B., Lotspeich, L., Kwon, H., Buonocore, M., Lammers, C., Reiss, A., and Amaral, D. (2004). The amygdala is enlarged in children but not adolescents with autism; the hippocampus is enlarged at all ages. *J Neurosci.*, 24:6392–6401.

Scotina, A. and Gutman, R. (2019). Matching Algorithms for Causal Inference With Multiple Treatments. *Stat. Med.*, 38:3139–3167.

Sears, L., Vest, C., Mohamed, S., Bailey, J., Ranson, B., and Piven, J. (1999). An MRI study of the basal ganglia in autism. *Prog Neuropsychopharmacol Biol Psychiatry*, 23:613–624.

Seeley, W., Menon, V., Schatzberg, A., Keller, J., Glover, G., Kenna, H., Reiss, A., and Greicius, M. (2007). Dissociable intrinsic connectivity networks for salience processing and executive control. *J. Neurosci*, 27:2349–2356.

Seidlitz, J., Váša, F., Shinn, M., Romero-Garcia, R., Whitaker, K., Vértes, P., Wagstyl, K., Kirkpatrick Reardon, P., Clasen, L., Liu, S., Messinger, A., Leopold, D., Fonagy, P., Dolan, R., Jones, P., Goodyer, I., Consortium, N., Raznahan, A., and Bullmore, E. (2018). Morphometric Similarity Networks Detect Microscale Cortical Organization and Predict Inter-Individual Cognitive Variation. *Neuron*, 97:231–247.e7.

Selvaraju, R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *2017 IEEE International Conference on Computer Vision (ICCV)*.

Sergerie, K., Chochol, C., and Armony, J. (2008). The role of the amygdala in emotional processing: a quantitative meta-analysis of functional neuroimaging studies. *Neuroscience and Biobehavioral Reviews*, 32:811–830.

Serrano, M., Maguitman, A., Boguña, M., Fortunato, S., and Vespignani, A. (2007). Decoding the structure of the www: A comparative analysis of web crawls. *ACM Transactions on the Web*, 1.

Sharda, M., Foster, N., Tryfon, A., Doyle-Thomas, K., Ouimet, T., Anagnostou, E., Evans, A., Zwaigenbaum, L., Lerch, J., Lewis, J., Hyde, K., and Group., N. A. I. (2017). Language ability predicts cortical structure and covariance in boys with Autism Spectrum Disorder. *Cerebral Cortex*, 27:1849–1862.

Shen, D., Wu, G., and Suk, H. (2017a). Deep Learning in Medical Image Analysis. *Annu Rev Biomed Eng.*, 19:221–248.

Shen, M. and Piven, J. (2017). Brain and behavior development in autism from birth through infancy. *Dialogues Clin Neurosci*, 19:325–333.

Shen, X., Reus, L., Cox, S., Adams, M., Liewald, D., Bastin, M., Smith, D., Deary, I., Whalley, H., and McIntosh, A. (2017b). Subcortical volume and white matter integrity abnormalities in major depressive disorder: findings from UK Biobank imaging data. *Sci Rep*, 7.

Shin, H., Roth, H., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., and Summers, R. (2016). Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Trans Med Imaging*, 35:1285–1298.

Shulman, G. L., Fiez, J. A., Corbetta, M., Buckner, R., Miezin, F., Raichle, M., and Petersen, S. (1997). Common Blood Flow Changes across Visual Tasks: II. Decreases in Cerebral Cortex. *J. Cognit. Neurosci.*, 9:648–663.

Simas, T. (2012). *Stochastic Models and Transitivity in Complex Networks*. PhD dissertation, Indiana University.

Simas, T., Chattopadhyay, S., Hagan, C., Kundu, P., Patel, A., Holt, R., Floris, D., Graham, J., Ooi, C., Tait, R., Spencer, M., Baron-Cohen, S., Sahakian, B., Bullmore, E., Goodyer, I., and Suckling, J. (2015a). Semi-Metric Topology of the Human Connectome: Sensitivity and Specificity to Autism and Major Depressive Disorder. *PLoS One*, 10.

Simas, T., Chattopadhyay, S., Hagan, C., Kundu, P., Patel, A., Holt, R., Floris, D., Graham, J., Ooi, C., Tait, R., Spencer, M., Baron-Cohen, S., Sahakian, B., Bullmore, E., Goodyer, I., and Suckling, J. (2015b). Semi-Metric Topology of the Human Connectome: Sensitivity and Specificity to Autism and Major Depressive Disorder. *PLoS One*, 10.

Simas, T. and Rocha, L. (2014). Distance Closures on Complex Networks. *arXiv*.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2014). Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In *Workshop at International Conference on Learning Representations*.

Simpson, J., Snyder, A., Gusnard, D., and Raichle, M. (2001). Emotion-induced changes in human medial prefrontal cortex: I. During cognitive task performance. *PNAS*, 98:683–687.

Singh, M. and Gotlib, I. (2014). The Neuroscience of Depression: Implications for Assessment and Intervention. *Behav Res Ther.*, 62:60–73.

Singh, S. and Singh, N. (2017). Object classification to analyze medical imaging data using deep learning. *International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), Coimbatore*, page 1–4.

Skidmore, F., Korenkevych, D., Liu, Y., He, G., Bullmore, E., and Pardalos, P. (2011). Connectivity brain networks based on wavelet correlation analysis in Parkinson fMRI data Author links open overlay panel. *Neuroscience Letters*, 499:47–51.

Skudlarski, P., Jagannathan, K., Calhoun, V., Hampson, M., Skudlarska, B., and Pearlson, G. (2008). Measuring brain connectivity: diffusion tensor imaging validates resting state temporal correlations. *NeuroImage*, 43:554–561.

Smith, S., Fox, P., Miller, K., Glahn, D., Fox, P., Mackay, C., Filippini, N., Watkins, K., Toro, R., Laird, A., and Beckmann, C. (2009). Correspondence of the brain's functional architecture during activation and rest. *Proc. Natl. Acad. Sci.*, 106:13040–13045.

Smith, S., Miller, K., Salimi-Khorshidi, G., Webster, M., Beckmann, C., Nichols, T., Ramsey, J., and Woolrich, M. (2011). Network modelling methods for FMRI. *NeuroImage*, 54:875–891.

Smith, S. and Nichols, T. (2018). Statistical Challenges in "Big Data" Human Neuroimaging. *Neuron*, 97:263–268.

Smyser, C., Inder, T., Shimony, J., Hill, J., Degnan, A., Snyder, A., and Neil, J. (2010). Longitudinal Analysis of Neural Network Development in Preterm Infants. *Cerebral Cortex*, 20:2852–2862.

Sparks, B., Friedman, S., Shaw, D., Aylward, E., Echelard, D., Artru, A., Maravilla, K., Giedd, J., Munson, J., Dawson, G., and Dager, S. (2002). Brain structural abnormalities in young children with autism spectrum disorder. *Neurology*, 59:184–192.

Sporns, O. (2006). Small-world connectivity, motif composition, and complexity of fractal neuronal connections. *BioSystems*, 85:55–64.

Sporns, O. (2010). *Networks of the Brain*. The MIT Press, Cambridge, MA, 1st edition.

Sporns, O. and Betzel, R. (2016). Modular Brain Networks. *Annu Rev Psychol.*, 67:613–640.

Sporns, O., Honey, C., and K otter, R. (2007). Identification and classification of hubs in brain networks. *PLoS One*, 2:e1049.

Sporns, O., Tononi, G., and Edelman, G. (2000). Theoretical Neuroanatomy: Relating Anatomical and Functional Connectivity in Graphs and Cortical Connection Matrices. *Cerebral Cortex*, 10:127–141.

Sporns, O. and Zwi, J. (2004). The small world of the cerebral cortex. *Neuroinformatics*, 2:145–162.

Sridharan, D., Levitin, D., and Menon, V. (2008). A critical role for the right fronto-insular cortex in switching between central-executive and default-mode networks. *PNAS*, 105:12569–12574.

Stanfield, A., McIntosh, A., Spencer, M., Philip, R., Gaur, S., and Lawrie, S. (2008). Towards a neuroanatomy of autism: a systematic review and meta-analysis of structural magnetic resonance imaging studies. *Eur Psychiatry*, 23:289–99.

Stanley, M., Moussa, M., Paolini, B., Lyday, R., Burdette, J., and Laurienti, P. (2013). Defining nodes in complex brain networks. *Front. Comput. Neurosci.*, 7:169.

Stephan, K., Hilgetag, C., Burns, G., O'Neill, M., Young, M., and K otter, R. (2000). Computational analysis of functional connectivity between areas of primate cerebral cortex. *Philos Trans R Soc Lond B Biol Sci*, 355:111–126.

Stevens, J. and Hamann, S. (2012). Sex differences in brain activation to emotional stimuli: a meta-analysis of neuroimaging studies. *Neuropsychologia*, 50:1578–1593.

Stuart, E. (2010). Matching methods for causal inference: A review and a look forward. *Stat. Sci.*, 25:1–21.

Su, R., Rounds, J., and Armstrong, P. (2009). Men and things, women and people: a meta-analysis of sex differences in interests. *Psychol Bull.*, 135:859–884.

Subbaraju, V., Suresh, M., Sundaram, S., and Narasimhan, S. (2017). Identifying differences in brain activities and an accurate detection of autism spectrum disorder using resting state functional-magnetic resonance imaging: A spatial filtering approach. *Med Image Anal*, 35:375–389.

Suckling, J., Simas, T., Chattopadhyay, S., Tait, R., Su, L., Williams, G., Rowe, J., and O'Brien, J. (2015). A Winding Road Alzheimers Disease Increases Circuitous Functional Connectivity Pathways. *Frontiers in Computational Neuroscience*, 9.

Sung, Y.-W., Kawachi, Y., Choi, U.-S., Kang, D., Abe1, C., Otomo, Y., and Ogawa, S. (2018). A Set of Functional Brain Networks for the Comprehensive Evaluation of Human Characteristics. *Front. Neurosci.*, 12:149.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2014). Intriguing properties of neural networks. *ICLR 2013*.

Tajbakhsh, N., Gurudu, S., and Liang, J. (2015). A comprehensive computer-aided polyp detection system for colonoscopy videos. *Information Processing in Medical Imaging*, 24:327–338.

Tajbakhsh, N. and Liang, J. (2015). Computer-aided pulmonary embolism detection using a novel vessel-aligned multi-planar image representation and convolutional neural networks. *Medical Image Computing and Computer-Assisted Intervention MICCAI*.

Takahashi, H., Matsuura, M., Yahata, N., Koeda, M., Suhara, T., and Okubo, Y. (2006). Men and women show distinct brain activations during imagery of sexual and emotional infidelity. *Neuroimage*, 32:1299–1307.

Tejwani, R., Liska, A., You, H., Reinen, J., and Das, P. (2017). Autism Classification Using Brain Functional Connectivity Dynamics and Machine Learning. *ArXiv*.

Thompson, W. and Fransson, P. (2015). The frequency dimension of fMRI dynamic connectivity: Network connectivity, functional hubs and integration in the resting brain. *NeuroImage*, 121:227–242.

Tijms, B., Seriès, P., Willshaw, D., and Lawrie, S. (2012). Similarity-based extraction of individual networks from gray matter MRI scans. *Cereb. Cortex*, 22:1530–1541.

Tixier, A., Nikolentzos, G., Meladianos, P., and Vazirgiannis, M. (2017). Classifying Graphs as Images with Convolutional Neural Networks. *arXiv*.

Tomasi, D. and Volkow, N. (2010). Functional connectivity density mapping. *Proc. Natl. Acad. Sci.*, 107:9885–9890.

Tomasi, D. and Volkow, N. (2011a). Association between functional connectivity hubs and brain networks. *Cereb Cortex*, 21:2003–2013.

Tomasi, D. and Volkow, N. (2011b). Gender differences in brain functional connectivity density. *Human Brain Mapping*, 33:849–860.

Traag, V. and Bruggeman, J. (2009). Community detection in networks with positive and negative links. *Phys. Rev. E.*, 80:036115.

Turner, K., Frost, L., Linsenbardt, D., McIlroy, J., and M uller, R. (2006). Atypically diffuse functional connectivity between caudate nuclei and cerebral cortex in autism. *Behav Brain Funct*, 2.

Tyszka, J., Kennedy, D., Adolphs, R., and Paul, L. (2011). Intact bilateral resting-state networks in the absence of the corpus callosum. *J Neurosci*, 31:15154–15162.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., and Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, 15:273–289.

Uddin, L., Mooshagian, E., Zaidel, E., Scheres, A., Margulies, D., Clare Kelly, A., Shehzad, Z., Adelstein, J., Castellanos, F., Biswal, B., and Milham, M. (2008). Residual functional connectivity in the split-brain revealed with resting-state fMRI. *Neuroreport*, 19:703–709.

Uematsu, A., Matsui, M., Tanaka, C., Takahashi, T., Noguchi, K., Suzuki, M., and Nishijo, H. (2012). Developmental trajectories of amygdala and hippocampus from infancy to early adulthood in healthy individuals. *PLoS One*, 7:e46970.

van den Heuvel, M., Kahn, R., Goñi, J., and Sporns, O. (2012). High-cost, high-capacity backbone for global brain communication. *PNAS*, 109:11372–11377.

van den Heuvel, M. and Sporns, O. (2011). Rich-Club Organization of the Human Connectome. *Journal of Neuroscience*, 31:15775–15786.

van den Heuvel, M. and Sporns, O. (2013). Network hubs in the human brain. *Cell*, 17:683–689.

van den Heuvel, M., Stam, C., Kahn, R., and Pol, H. (2009). Efficiency of functional brain networks and intellectual performance. *J Neurosci*, 29:7619–7624.

van Rooij, D., Anagnostou, E., Arango, C., Auzias, G., Behrmann, M., Busatto, G., Calderoni, S., Daly, E., Deruelle, C., Di Martino, A., Dinstein, I., Duran, F., Durston, S., Ecker, C., Fair, D., Fedor, J., Fitzgerald, J., Freitag, C., Gallagher, L., Gori, I., Haar, S., Hoekstra, L., Jahanshad, N., Jalbrzikowski, M., Janssen, J., Lerch, J., Luna, B., Martinho, M., McGrath, J., Muratori, F., Murphy, C., Murphy, D., O'Hearn, K., Oranje, B., Parellada, M., Retico, A., Rosa, P., Rubia, K., Shook, D., Taylor, M., Thompson, P., Tosetti, M., Wallace, G., Zhou, F., and Buitelaar, J. (2017). Cortical and Subcortical Brain Morphometry Differences Between Patients With Autism Spectrum Disorder and Healthy Individuals Across the Lifespan: Results From the ENIGMA ASD Working Group. *American Journal of Psychiatry*, 175:359–369.

Vincent, J., Patel, G., Fox, M., Snyder, A., Baker, J., Van Essen, D., Zempel, J., Snyder, L., Corbetta, M., and Raichle, M. (2007). Intrinsic functional architecture in the anaesthetized monkey brain. *Nature*, 447:83–86.

Vossel, S., Geng, J., and Fink, G. (2014). Dorsal and Ventral Attention Systems: Distinct Neural Circuits but Collaborative Roles. *Neuroscientist*, 20:150–159.

Wang, J., Wang, L., Zang, Y., Yang, H., Tang, H., Gong, Q., Chen, Z., Zhu, C., and He, Y. (2009). Parcellation-dependent small-world brain functional networks: A resting-state fMRI study. *Human Brain Mapping*, 30:1511–1523.

Wang, W., Liu, J., Shi, S., Liu, T., Ma, L., Ma, X., Tian, J., Gong, Q., and Wang, M. (2018). Altered Resting-State Functional Activity in Patients With Autism Spectrum Disorder: A Quantitative Meta-Analysis. *Front. Neurol.*, 9:556.

Wang, X., Ren, Y., and Zhang, W. (2017). Depression Disorder Classification of fMRI Data Using Sparse Low-Rank Functional Brain Network and Graph-Based Features. *Computational and Mathematical Methods in Medicine*.

Wang, Z., Dai, Z., Gong, G., Zhou, C., and He, Y. (2015). Understanding structural-functional relationships in the human brain: a large-scale network perspective. *The Neuroscientist*, 21:290–305.

Watts, D. and Strogatz, S. (1998). Collective dynamics of 'small-world' networks. *Letters to Nature*, 393:440–442.

Weinstein, M., Ben-Sira, L., Levy, Y., Zachor, D., Itzhak, E., Artzi, M., Tarrasch, R., Eksteine, P., Hendler, T., and Bashat, D. (2011). Abnormal white matter integrity in young children with autism. *Hum. Brain Mapp.*, 32:534–543.

Weissman-Fogel, I., Moayedi, M., Taylor, K., Pope, G., and Davis, K. (2010). Cognitive and default-mode resting state networks: do male and female brains 'rest' differently? *Hum Brain Mapp*, 31:1713–1726.

Weng, S., Wiggins, J., Peltier, S., Carrasco, M., Risi, S., Lord, C., and Monk, C. (2010). Alterations of resting state functional connectivity in the default network in adolescents with autism spectrum disorders. *Brain Res*, 1313:202–214.

Whitwell, J. (2009). Voxel-Based Morphometry: An Automated Technique for Assessing Structural Changes in the Brain. *J Neurosci.*, 29:9661–9664.

Wierenga, L., Sexton, J., Laake, P., Giedd, J., and Tamnes, C. (2017). A key characteristic of sex differences in the developing brain: greater variability in brain structure of boys than girls. *Cereb Cortex*, 28:2741–2751.

Wolff, J., Jacob, S., and Elison, J. (2018). The journey to autism: Insights from neuroimaging studies of infants and toddlers. *Dev Psychopathol*, 30:479–495.

Wrase, J., Klein, S., Gruesser, S., et al. (2003). Gender differences in the processing of standardized emotional visual stimuli in humans: a functional magnetic resonance imaging study. *Neuroscience Letters*, 348:41–45.

Wright, I., McGuire, P., Poline, J., Travere, J., Murray, R., Frith, C., Frackowiak, R., and Friston, K. (1995). A voxel-based method for the statistical analysis of gray and white matter density applied to schizophrenia. *Neuroimage*, 2:244–252.

Wright, I., Sharma, T., Ellison, Z., McGuire, P., Friston, K., Brammer, M., Murray, R., and Bullmore, E. (1999). Supra-regional brain systems and the neuropathology of schizophrenia. *Cereb. Cortex*, 9:366–378.

Wu, B., Chen, W., Fan, Y., Zhang, Y., Hou, J., Liu, J., and Zhang, T. (2019). Tencent ML-Images: A Large-Scale Multi-Label Image Database for Visual Representation Learning. *IEEE Access*, 7:172683–172693.

Xin, J., Zhang, Y., Tang, Y., and Yang, Y. (2019). Brain Differences Between Men and Women: Evidence From Deep Learning. *Front. Neurosci.*

Yan, C., Gong, G., Wang, J., Wang, D., Liu, D., Zhu, C., and He, Y. (2011). Sex- and brain size–related small-world structural cortical networks in young adults: a DTI tractography study. *Cereb Cortex*, 21:449–458.

Yang, J. and Hofmann, J. (2015). Action observation and imitation in autism spectrum disorders: an ALE meta-analysis of fMRI studies. *Brain Imaging and Behavior*, 10:960–969.

Yao, Z., Hu, B., Xie, Y., Moore, P., and Zheng, J. (2015). A review of structural and functional brain networks: small world and atlas. *Brain Informatics*, 2:45–52.

Yen, J. (1971). Finding the K Shortest Loopless Paths in a Network. *Management Science*, 17:712–716.

Yoldemir, B., Ng, B., and Abugharbieh, R. (2015). Coupled Stable Overlapping Replicator Dynamics for Multimodal Brain Subnetwork Identification. *IPMI 2015. Lecture Notes in Computer Science*, 9123.

Yoo, S., Han, C., Shin, J., Seo, S., Na, D., Kaiser, M., and Seong, J. (2015). A network flow-based analysis of cognitive reserve in normal ageing and Alzheimer's Disease. *Scientific reports*, 5:10057.

Yoshida, K., Shimizu, Y., Yoshimoto, J., Takamura, M., Okada, G., Okamoto, Y., Yamawaki, S., and Doya, K. (2017). Prediction of clinical depression scores and detection of changes in whole-brain using resting-state functional MRI data with partial least squares regression. *PLoS ONE*, 12:e0179638.

Zalesky, A., Fornito, A., Harding, I., Cocchi, L., Y ucel, M., Pantelis, C., and Bullmore, E. (2010). Whole-brain anatomical networks: does the choice of nodes matter? *NeuroImage*, 50:970–983.

Zeiler, M. (2012). ADADELTA: An Adaptive Learning Rate Method. *arXiv*.

Zeiler, M. and Fergus, R. (2013). Visualizing and Understanding Convolutional Networks. *Computer Vision – ECCV 2014. Lecture Notes in Computer Science*, 8689.

Zeiler, M., Krishnan, D., Taylor, G., and Fergus, R. (2010). Deconvolutional networks. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Zeng, L., Shen, H., Liu, L., Wang, L., Li, B., Fang, P., Zhou, Z., Li, Y., and Hu, D. (2012). Identifying major depression using whole-brain functional connectivity: a multivariate pattern analysis. *Brain*, 135:1498–1507.

Zhan, L., Jenkins, L., Wolfson, O., GadElkarim, J., Nocito, K., Thompson, P., Ajilore, O., Chung, M., and Leow, A. (2017). The Significance of Negative Correlations in Brain Connectivity. *J Comp Neurol*, 525:3251–3265.

Zhang, J., Wang, J., Wu, Q., Kuang, W., Huang, X., and Gong, Q. (2011a). Disrupted Brain Connectivity Networks in Drug-Naive, First-Episode Major Depressive Disorder. *Biological Psychiatry*, 70:334–342.

Zhang, S., Li, X., Lv, J., Jiang, X., Guo, L., and Liu, T. (2016a). Characterizing and Differentiating Task-based and Resting State FMRI Signals via Two-stage Sparse Representations. *Brain Imaging Behav.*, 10:21–32.

Zhang, W., Groen, W., Mennes, M., Greven, C., Buitelaar, J., and Rommelse, N. (2018). Revisiting subcortical brain volume correlates of autism in the ABIDE dataset: effects of age and sex. *Psychological Medicine*, 48:654–668.

Zhang, W., K. Doi, K., Giger, M., Wu, Y., Nishikawa, R., and Schmidt, R. (1994). Computerized detection of clustered microcalcifications in digital mammograms using a shift-invariant artificial neural network. *Medical Physics*, 21:517–524.

Zhang, Z., Liao, W., Zuo, X.-N., Wang, Z., Yuan, C., Jiao, Q., Chen, H., Biswal, B., Lu, G., and Liu, Y. (2011b). Resting-state brain organization revealed by functional covariance networks. *PLoS ONE*, 6:e28817.

Zhang, Z., Telesford, Q., Giusti, C., Lim, K., and Bassett, D. (2016b). Choosing Wavelet Methods, Filters, and Lengths for Functional Brain Network Construction. *PLoS ONE*, 11:e0157243.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2015a). Learning Deep Features for Discriminative Localization. *CVPR'16*.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2015b). Object Detectors Emerge in Deep Scene CNNs. *International Conference on Learning Representations*.

Zuo, X., Ehmke, R., Mennes, M., Imperati, D., Castellanos, F., Sporns, O., and Milham, M. (2011). Network Centrality in the Human Functional Connectome. *Cerebral Cortex*, 22:1862–1875.